

统计学系列

统计学的世界

让你学会用手中的少量数据，
对重大问题做出明智的决策。

第五版
Fifth Edition

STATISTICS
Concepts and Controversies

[美] 戴维·S·穆尔 著
(David S. Moore)



中信出版社
CITIC PUBLISHING HOUSE

这不是一本谈统计理论的书。本书主要谈的是对统计概念的应用，以及统计概念对日常生活、公共政策和许多其他领域研究的影响。书中没有繁琐的计算，只要看得懂而且会用简单的方程式就够了。不过我在这里要提醒读者：这本书的重点在思考，而思考要比套用数学公式更能训练思维。

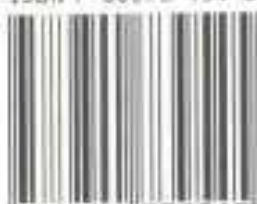
本书把统计概念分成四部分呈现：

- 数据的产生：数据怎么得来，非常重要，这是统计当中影响最大的概念。
- 资料分析：你会学到，即使用很简单的方法，也能很睿智地解读数据。
- 概率：利用概率进行思考，可以帮你把事实和无关紧要的干扰信息分离。
- 统计推论：让你学会用手中的少量数据，对一个较大的总体做出结论。

许多统计学家在第二次世界大战中发挥了重大的作用。沃德是其中之一。他发明的一些统计方法，在战时被视为军事机密。沃德在被咨询飞机上什么部位的钢板需要加强时，画了飞机的轮廓，并且标出返航的战斗机上受敌军创伤的弹孔位置。资料积累一段时间后，机身各部位几乎都被填满了。于是沃德建议，把剩下少数几个没有弹孔的位置加强，因为这些部位被击中的飞机都没有返航。

——摘自原书

ISBN 7-80073-954-6



9 787800 739545 >

www.publish.citic.com



ISBN 7-80073-954-6/F · 600

定价：79.00元

统计学的世界 concepts and controversies

【页 数】 633 页

【作 者】 (美) 戴维·S·穆尔著 郑惟厚译

【丛书名】 统计学系列

【形态项】 633 页 ; 26cm

【读秀号】 000003048473

【出版项】 中信出版社 , 2003

【ISBN 号】 7-80073-954-6 / C8

【原书定价】 CNY79.00 网上购买

【主题词】 统计学(学科: 基本知识)

【参考文献格式】(美) 戴维·S·穆尔著 郑惟厚译. 统计学的世界 concepts and controversies. 中信出版社, 2003.

内容提要:

统计学的思想和各种统计数据对政府、社会乃至我们的工作和日常生活都产生着直接的影响, 这种影响可能远远超乎你的想像。通过阅读本书, 你将对我们这个世界有一个更完整、更清晰的认识。本书一点儿也不枯燥乏味, 恰恰相反, 它是那样生动有趣, 深入浅出地把统计学的概念和分析方法呈现在你面前。通过一个个真实的小故事, 本书能让你在会心的微笑中不知不觉地增长专业知识, 提高分析水平。这是一本能给你带来乐趣的书, 也是一本能让你更加睿智的书。

作者简介:

戴维·S 穆尔, 美国普度大学统计学教授。资料分析法的创始人之一, 也是美国统计学会/美国数学学会委员会委员。穆尔教授提介统计教学法的改革, 并在数学方面取得了卓越的成就, 曾获美国数学学会 1995 年杰出教学奖。

目录:

写给教师 统计可以当作公共课程来教

前言 什么是统计

统计与你 这本书里谈些什么?

第一部分 产生数据

第一章 数据从何而来?

第二章 好样本和坏样本

第三章 样本告诉我们什么?

第四章 真实世界中的抽样调查

第五章 实验面面观

第六章 真实世界中的实验

第七章 数据伦理
第八章 度量
第九章 数字合不合理？
第一部分 复习

第二部分 整合数据

第十章 好的图及坏的图
第十一章 用图形呈现分布
第十二章 用数字描述分布
第十三章 正态分布
第十四章 描述相关关系的方法：散布图和相关系数
第十五章 描述相关关系：回归、预测及因果关系
第十六章 消费者物价指数和政府统计
第二部分 复习

第三部分 机遇

第十七章 考虑可能性
第十八章 概率模型
第十九章 模拟
第二十章 赌场的优势：期望值
第三部分 复习

第四部分 推论

第二十一章 什么是置信区间
第二十二章 什么是显著性检验
第二十三章 统计推论的使用与滥用
第二十四章 双向表及卡方检验
第二十五章 有关总体平均数的推论
第四部分 复习

注释与资料出处

部分习题解答

表 A 随机数字

表 B 正态分布的百分位数

编辑推荐：

这不是一本谈统计理论的书。本书主要谈的是对统计概念的应用，以及统计概念对日常生活、公共政策和许多其他领域研究的影响。书中没有繁锁的计算，只要看得懂而且会用简单的方程式就够了。不过我在这里要提醒读者：这本书的重点在思考，而思考要比套用数学公式更能训练思维。

统计学系列

统计学的世界

让你学会用手中的少量数据，
对重大问题做出明智的决策。

第五版
Fifth Edition

STATISTICS
Concepts and Controversies

[美] 戴维·S·穆尔 著
郑惟厚 译

中信出版社
CITIC PUBLISHING HOUSE

图书在版编目 (CIP) 数据

统计学的世界/[美]穆尔著;郑惟厚译. —北京:中信出版社, 2003. 9

书名原文: Statistics: concepts and controversies

ISBN 7-80073-954-6

I. 统… II. ①穆… ②郑… III. 统计学 - 通俗读物 IV. C8-49

中国版本图书馆 CIP 数据核字 (2003) 第 084824 号

Statistics: concepts and controversies

© 2001, by W. H. Freeman and Company.

Simplified Chinese Translation copyright © 2003 by CITIC Publishing House

Published by arrangement with W. H. FREEMAN AND COMPANY, a division of BEDFORD.

FREEMAN AND W. H. FREEMAN AND COMPANY LLC.

ALL RIGHTS RESERVED.

本书译文由天下文化出版公司授权使用。

统计学的世界 (第五版)

TONGJIXUE DE SHIJIE

著 者: [美]戴维·S·穆尔

译 者: 郑惟厚

责任编辑: 谢楠

出版发行: 中信出版社 (北京市朝阳区东外大街亮马河南路 14 号塔园外交办公大楼 邮编 100600)

经 销 者: 中信联合发行有限公司

承 印 者: 中国电影出版社印刷厂

开 本: 787mm × 1092mm 1/16 印 张: 41.5 字 数: 543 千字

版 次: 2003 年 11 月第 1 版 印 次: 2003 年 11 月第 1 次印刷

京权图字: 01-2003-5930

书 号: ISBN 7-80073-954-6/F · 600

定 价: 79.00 元

版权所有·侵权必究

凡购本社图书,如有缺页、倒页、脱页,由发行公司负责退换 服务热线:010-85322521

E-mail: sales @ citicpub.com

010-85322522

写给教师

统计可以当作公共课程来教

《统计学的世界》是一本公共课程的统计(statistics)书，也就是说，是非数学系学生公共教育的一部分。这本书的内容，是从我在给普度大学(Purdue University)文学院一、二年级学生开课的经验中，累积发展出来的。很高兴的是其他许多教师也觉得，这本书对各领域的学生都有用，甚至对哲学系和药系的学生也是如此。这本第五版经过了大幅度的修订，差不多可以算是本新书。不过目标仍和前几版相同，就是要把统计以不同的面貌呈现给大家，即不把统计当作专业的工具，而是当作受过教育的人应有的文化素养。

统计和文科的关系

统计在一般人的心目中，是最不“文科”的学科了。要是有人说统计的好话，多半是在说它的用处。医疗领域的专家必须有统计知识，才能了解医学研究报告的内容；经理人需要统计，因为大量的数字需要经过有效浓缩，才能看得出所以然；老百姓需要了解统计，才能知道民意调查和消费者物价指数(CPI, Consumer Price Index)到底是怎么回事。因为数据(data)和机遇(chance)无所不在，所以我们的广告词可以这样写：每个人都用得到统计，甚至因统计而获利多多。

以上说的是事实。我还可以进一步地说，对于大多数的学生来讲，这本书所持的强调“观念”，以及用言语来说明的方式，帮助学生为将来学习统计所打下的基础，要优于一般以“方法”为主的介绍方式。美国统计协会(American Statistical Association)和美国数学学会(Mathematical Association of American)的联合课程



委员会就曾建议，任何统计的入门课程，都应该“强调如何做统计思考”而且内容应该“多一些数据和观念，少一点公式和推导过程”。这本书就像这项建议所说的，它的内涵符合素质教育的精神，也就是有许多观念、许多思考，而只有简单的数据、少量的公式，且没有任何正式推导。

应该学习统计观念的另一个理由是：统计其实属于文科领域。人文教育强调基本的思维能力，也就是可以应用在各种情境的一般探索方法。传统的人文学科中呈现了这些方法，比如文学和历史研究、人类社会的政治分析及社会分析、探讨大自然的实验科学，以及数学领域中抽象概念及演绎的力量。统计可归类于文科领域，因为根据不确定的经验数据(empirical data)做推论，用的也是类似的一般思维方法。这本书的两个主题——数据及机遇，普遍存在于我们的日常经验当中。虽然我们数学工具来处理数据和机遇问题，但数学所提供的观念，却不是纯然数学的。事实上，心理学家已经提出具说服力的论点，认为若想要加强我们对日常生活中所遇到的数据或机遇现象做出合理推论的能力，精通数学并没有多大帮助。

这本书在作者的能力范围和读者可接受的限度下，提出的观点是：统计是一种独立且基本的思考方法。书的重点是在统计思考，也有人称为认识数(quantitative literacy)*。

*译注：literacy 指识字，
quantitative literacy 就指认识
数。

本书的类型

统计学的书有的专讲理论，有的专讲方法，但这本书两者都不是。《统计学的世界》讲的是统计概念、统计推理，还有统计与公共政策和包括从药学到社会学等人文科学的关系。我加入了许多基本的绘图及数值技巧，来落实概念，支持推理。学生通过处理数据，可以学到怎么样来看数据。不过我可没有让技巧超越观念变成主角。我打算用言语来教而不是用代数来教，鼓励讨论甚至争辩，而不是只教计算，当然有一些计算是必须的。而从目录就可以看得出，本书的涵盖范围比传统的统计教科书广得多。以素质教育的精神来看，我宁广不宁精。

虽然本书的内容不那么正经八百，但它仍是一本教科书。书的编排是针对有系统的学习所规划，书中有许多习题，而且其中



有不少是要求学生讨论或做判断。即使对于那些主动阅读本书且乐在其中的同学来说，也应该做做习题而不是只读正文。而教师们请小心，虽然本书所用到的数学层次很低，但它并不是一本轻松的教科书。本书重点放在观念和推理上，教起来反而会比满篇都是公式的教科书还辛苦。

前言

什么是统计

统计是用来处理数据的。数据由数字组成，但它不仅是单纯的数字而已。数据是有内容的数字。比如说，光是 10.5 这个数字本身并没什么含意，但是假如我们得知，一个朋友的新生儿出生时重 10.5 磅，我们会恭喜她生了个健康宝宝。我们根据数字，配合上下文并和常识衔接，就可以做出判断。我们知道 10.5 磅重的婴儿相当大，而且也知道婴儿不太可能重 10.5 英两或 10.5 公斤。数字加上上下文才能提供有效的信息。

统计是从数据中找出信息，并且做出结论。我们用的工具是图表和计算，加上常识判断。我们先来简单地快速看一下，媒体和热门政治或社会议题如何处理数据和统计研究，借此展开讨论统计的序幕。稍后我们会对这里提到的例子做更详尽的审视。

数据胜过轶闻

信仰不能取代数字

斯潘塞 (Henry Spencer)

轶闻引人注目，是因为它很突出，所以会深入人心。轶闻能使议题人性化，所以新闻报道常常以轶闻当开场或结尾。但它并不足以当成做决定的根据，而且正因为它们很突出，所以反而常常会产生误导。应该注意的是，声明的背后有没有数据支持，而不是有没有动人的故事。

住在高压电线附近会导致儿童得白血病吗？美国国家癌症研究所 (National Cancer Institute) 花费 5 年的时间和 500 万美元为这个问题搜集资料。结论是：在白血病和暴露在高压电线所产生的磁



场之间，找不到相关关系。在《新英格兰医学期刊》(*New England Journal of Medicine*)上和这篇研究报告同时登出的评论中严厉提出，“应该马上停止浪费我们的研究资源”在这个问题上。

现在试着比较以下两者的影响：一是电视上对于一项耗时5年，花费500万美元的调查结果的新闻报道，另一是电视访问一位能说会道的母亲，她的孩子得了白血病，而且他们恰巧住在高压电线附近。在大众心目中，每次都是轶闻得胜。但我们应该要心存疑问。数据比轶闻可靠，因为数据可以有系统地描绘出整体的情况，而轶闻只聚焦于少数特例。

我忍不住要加上一句“数据胜过自封的专家”。大部分媒体心目中的平衡报道，就是指在不同立场的两方面，各找一位“专家”来发表简短的评论。我们永远也没法知道，是否某位专家表达的是真正具有代表性的意见，而另一位只是照顾私人利益的骗子。而且由于媒体对于制造冲突的喜好，社会大众现在以为，对应每一位专家，必有立场相反的另一位专家。如果你真正关心某个议题，应该试着判断数据透露何种信息，以及数据的品质如何。当然！的确有许多议题悬而未决，但也有许多议题只在不重视证据的人心中才悬而未决。你至少可以了解一下这些“专家”的背景，还有他们所引用的研究结果，是否曾刊载在审编制度严谨的期刊上。

数据从何而来非常重要

数字不会说谎，但说谎的人会想出办法。

格罗夫纳(Charles Grosvenor)

数据是数字，而数字总显得牢不可破。其实有的如此，有的却不是。任何统计研究中最重要的事，就是数据的来源。当专栏作者蓝德丝(Ann Landers)问她的读者：如果重新来过的话是否仍要生孩子？而回答的人当中有70%坚决说“不要”的时候，你对于蓝德丝引述自泪迹斑斑的信中，投书人泣诉他们的孩子如何像野兽一样时，不妨一笑置之。蓝德丝从事的是娱乐业，她邀请读者回答这个问题时，回应最热烈的应该是后悔有了孩子的父母。大部分的父母并不后悔有小孩。我们知道这个事实，是因为曾经



有人对许多父母做过意见调查，而且为了不偏向任何一个答案，访问的父母是随机(at random)抽取的。民意调查当然不是没有瑕疵，这点我们以后会谈到，但是比起邀请满肚子气的人来回答问题，民意调查显然高明多了。

即使是信誉卓著的期刊，也不一定能对坏数据免疫。《美国医学会期刊》(*Journal of the American Medical Association*)曾登过一篇文章，声称将冷却的液体经过管子打进胃里，可以缓解溃疡症状。病人的确对这种治疗有反应，但只是因为病人通常对自己信任的医师的权威信服，因此对他给的任何治疗都有反应。也就是说安慰剂*(placebo)发生了作用。后来终于有人起疑，做了有适当对照组(control)的研究，有些病人接受这项治疗，有些接受安慰剂，结果使用安慰剂这组的“表现”还稍好些呢。“没有比较，就没有结论”，是判断医学研究的一个很好的出发点。比如说，我对于最近突然盛行的“自然疗法”就有点存疑。这些疗法当中，极少会经过比较试验(comparative trial)，来证明它们不仅仅是装在有漂亮植物图片的瓶子里卖的安慰剂而已。

*译注：没有实际治疗作用的假治疗

小心潜在变量

我的钱够我这辈子用了，只要我不买东西。

梅森(Jackie Mason)

有报道说，在美国，设有赌场的县犯罪率较高。有位大学老师说，在网上修课的学生，比在教室里修课的学生表现好。政府的报告强调，受教育多的民众，比起受教育少的民众，赚钱要多得多。最好不要太快做结论。要先问：“是不是有什么他们没有告诉我的，可以解释这些事？”

有赌场的县的确犯罪率较高，但是在城市或是较穷的县，犯罪率照样比较高。怎样的县会有赌场呢？这些县是否在设立赌场之前，就有高犯罪率了呢？网上修课的学生学得比较好，但是和在教室修课的学生比起来，这些学生年龄比较大，底子也比较好，因此表现好很正常。教育程度高的人的确钱赚得多。但受教育多的人和受教育少的人比起来，平均来说他们的父母受的教育较多，也比较有钱。他们在较好的环境长大，上较好的学校。这些有利



条件让他们得到更多的教育，而即使他们不受这么多教育，这些条件也可能帮他们赚很多钱。

以上提到的这些研究，都报告了两个变量之间的关联，并且引导我们推断出，其中一个变量影响了另一个变量的结论。“赌场会增加犯罪率”以及“想要有钱就多读点书”是他们传达的信息。这些信息有可能是正确的。但也说不定我们看到的关联，大部分可以由隐藏在背景中的一些变量解释，比如会接受赌场的县的本质，以及高教育程度者生来就具备的有利条件。好的统计研究会考虑许多的背景变量。这当然要有技巧，不过你至少可以找一找，看他们有没有做。

变异无所不在

当事实改变时，我就改变主意。您呢？

凯恩斯(John Maynard Keynes)

如果你的舌下温超过 37°C ，是不是就代表你在发烧呢？也许不是。每个人的“正常”体温都会有一些差异。你自己的体温也有变化，早上 6 点时稍高些，下午 6 点就较低。美国政府宣布上个月失业率上升 0.1 个百分点，而新开发的建筑项目下降了 3 个百分点。股市因此遽升或遽降。股市波动常常是不理性的。政府的数据是根据好的样本得到的，是好的估计，但并不是百分之百的事实。同样的调查再做一遍，结果可能会有一些差别。而和经济有关的事件原本就会上上下下，而影响它的因素包括气候、罢工事件、假日以及各式各样其他原因。

很多人都像股市一样，会对数据的一些微小的变动做过度反应，而事实上这些改变并不是实质的改变，根本微不足道。以下是全美国最大的市场研究公司其领导人尼尔森(Arthur Nielsen)的经验之谈：

太多商界人士都对所有白纸黑字印出来的数字都同样的信任。他们认为数字就代表“事实”，要他们用“概率”(probability)的观点来看事情有点困难。他们未能看清，数字只是一种简化的表示法，它描述一个范



围，这范围说明我们对潜在情况的实际了解有多少。

变异是无所不在的。个体之间有差别，而对同一个个体多度量几次，结果也会不一样，并且几乎每件事都会随着时间而不同。至于一些自封的专家对每天股市变化的深入分析，或者对一场其实是最后一秒才定输赢的球赛，却硬要归罪于输球的球队的能力或特征，你尽可以对这些评论嗤之以鼻。

结论并不是百分之百的

数学定律不能百分之百确实地用在现实生活里：能百分之百确实地用数学定律描述的，就不是现实生活。

爱因斯坦 (Albert Einstein)

因为变异无所不在，所以统计结论并不是绝对的。大部分中年妇女会定期做乳房 X 光摄影，以期能早期发现乳癌。但乳房摄影是否真的可以减低死于乳癌的风险？高品质的统计研究发现，对于 50—64 岁的妇女来说，乳房摄影可以减少 26% 的死亡率。但这个数字是该年龄层妇女的一个平均数字。因为个别差异到处都有，所以对于不同的妇女来说，结果可能大不相同。有些每年做乳房摄影的妇女死于乳癌，而有一辈子没做过乳房摄影的妇女，却活到 100 岁，最后是因为骑摩托车出车祸而死亡。

总结报告事实上说的是“乳房摄影可以将乳癌死亡率减少 26% [95% 置信区间 (confidence interval) 为 17%—34%]”。而 26% 这个数字，根据尼尔森的说明，是“某范围的简化说法，这范围描述我们对隐藏情况的实际了解有多少”。在此例当中，这个范围是 17%—34%，而我们有 95% 的信心，真正的比例会落在这个范围内。也就是说，我们相当有把握，但不完全确定。一旦你超越新闻报道的层次，可以找找看有没有诸如“95% 信心”及“有统计上的显著意义 (statistically significant)”等字眼，这代表这个研究的结果虽不能说百分之百确定，但已相当有把握。



数据可反映社会价值

要用统计骗人很容易。但是不用统计，骗人更容易。

莫斯提勒(Frederick Mosteller)

好的数据确实胜过轶闻。比起轶闻和光大声嚷嚷预测未来，数据要客观得多。统计和其他的公开论述比起来，它根据事实且较科学又较理性。对于争论性的议题，统计研究应该比其他大部分证据受到更多的重视。不过世界上没有百分之百客观这回事。社会环境会影响我们对于要度量什么及如何度量的决定，因此也就影响到统计。

以自杀率来说，在不同国家间就有很大的差别。各国公布的自杀率有差别，似乎大部分要归因于社会观念，而不是因为自杀率真的有差别。自杀人数的计算，是根据死亡证明书。填写死亡证明书(证明书上的细节，依不同的国家或不同的州而有别)的官员，对于比如说像没有目击证人的溺死或摔死事件，可以决定要追究的程度。在视自杀为耻辱的地方，就有较多的自杀被报道为意外死亡。比方说，在大部分人信奉天主教的国家，所公布的自杀率就比其他国家低。日本文化中，有遇到羞辱时就光荣自杀的传统，这种传统使人们较不视自杀为耻辱，就使得日本的自杀事件，被报道的比例较高。而在其他国家，可能因为社会价值观的改变，使得自杀事件的计数增加。一个愈来愈普遍的看法是，抑郁症应该被视为生理上的疾病，而不是个性上的弱点，而自杀是这种病的悲剧性结束，不是道德瑕疵。因此使得家属和医生更愿意把自杀列为死因。

在不像自杀这么敏感的问题上，社会价值观依然可能影响到数据。美国的失业率，是每个月由劳工统计局(Bureau of Labor Statistics)用一个很大且很专业的全国性样本算出来的。但是“失业”的定义是什么呢？它是说虽然你想工作，却没有工作，而且过去两周曾经积极找工作。如果你已有两周没在找工作，你就不算是失业，而只是“不属于劳动人口”。这样定义失业率，反映了我们加诸于工作的价值观。如果换个方式定义，可能会得到很不



一样的失业率。

我的意思，不是叫你不要相信失业率的数字、失业率的定义已经稳定了好一段时间，所以我们可以从它看出趋势。在各个国家之间，这个定义也大致吻合，所以我們也可以在国际间做比较。数据是由不受政治干扰的专业人士算出来的。失业率是重要且有用的信息。我要强调的是，并不是每件重要的事都可以只用数字表示，还有，把事情简化为数字的人，会受到各种压力的影响，不管有意识还是无意识。

统计学的世界

简 明 目 录

第一部分 产生数据	1
第 1 章 数据从何而来?	2
第 2 章 好样本和坏样本	18
第 3 章 样本告诉我们什么?	34
第 4 章 真实世界中的抽样调查	58
第 5 章 实验面面观	84
第 6 章 真实世界中的实验	106
第 7 章 数据伦理	128
第 8 章 度量	150
第 9 章 数字合不合理?	174
第一部分 复习	192
 第二部分 整合数据	 205
第 10 章 好的图及坏的图	206
第 11 章 用图形呈现分布	232
第 12 章 用数字描述分布	254
第 13 章 正态分布	282
第 14 章 描述相关关系的方法:散布图和相关系数	306
第 15 章 描述相关关系:回归、预测及因果关系	330
第 16 章 消费者物价指数和政府统计	358
第二部分 复习	382
 第三部分 机遇	 399
第 17 章 考虑可能性	400
第 18 章 概率模型	420



第 19 章	模拟	436
第 20 章	赌场的优势：期望值	454
第三部分	复习	472
第四部分 推论		483
第 21 章	什么是置信区间	484
第 22 章	什么是显著性检验	504
第 23 章	统计推论的使用与滥用	522
第 24 章	双向表及卡方检验*	540
第 25 章	有关总体平均数的推论*	562
第四部分	复习	578

*本章内容较深，可列为选读。

统计学的世界

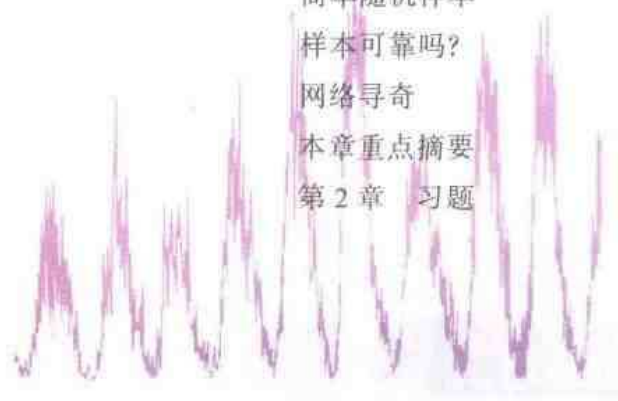
详细目录

写给教师	统计可以当作公共课程来教	I
前言	什么是统计	IV
统计与你	这本书里谈些什么?	XI

第一部分 产生数据 1

第1章	数据从何而来?	2
	来说说数据: 个体和变量	3
	观测研究	4
	抽样调查	6
	普查	9
	实验	10
	网络寻奇	12
	本章重点摘要	13
	第1章 习题	14

第2章	好样本和坏样本	18
	怎样可取得坏样本	19
	简单随机样本	21
	样本可靠吗?	25
	网络寻奇	26
	本章重点摘要	27
	第2章 习题	28





第3章 样本告诉我们什么?	34
从样本到总体	35
抽样变异	36
误差界限	40
置信叙述	43
从大总体抽样	45
统计学上的争议：选举民意调查该禁止吗？	46
网络寻奇	47
本章重点摘要	48
第3章 习题	49
第4章 真实世界中的抽样调查	58
抽样调查怎样出错	59
抽样误差	60
非抽样误差	62
问题的措辞	65
如何应对非抽样误差	66
真实世界中的抽样设计	67
相信调查结果之前该问的问题	71
网络寻奇	72
本章重点摘要	73
第4章 习题	74
第5章 实验面面观	84
谈谈实验	85
怎么样做坏实验	86
随机化比较实验	89
实验设计的逻辑	91
统计显著性	93
只能观测的时候怎么办	93
网络寻奇	96
本章重点摘要	97
第5章 习题	98



第6章 真实世界中的实验	106
--视同仁	107
双盲实验	108
统计学上的争议：到底是不是安慰剂？	110
拒绝参加、不合作者及退出者	111
我们的结论可以推广吗？	112
真实世界中的实验设计	116
配对及区集设计	117
网络寻奇	120
本章重点摘要	121
第6章 习题	122

第7章 数据伦理	128
首要原则	129
试验审查委员会	130
知情且同意	131
保密原则	132
临床试验	134
统计学上的争议：可以买希望？	137
行为及社会科学实验	138
网络寻奇	140
本章重点摘要	141
第7章 习题	142

第8章 度量	150
度量的基本原理	151
了解你的变量	153
有效量度和无效量度	154
准确和不准确量度	158
统计学上的争议：SAT 测验及大学入学申请	160
增加可靠程度，减少偏差	161
请同情可怜的心理学家	164



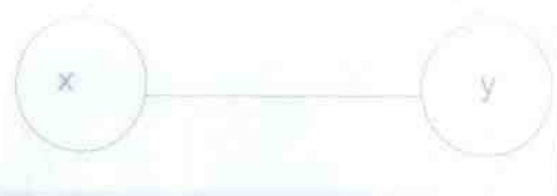


网络寻奇	166
本章重点摘要	167
第 8 章 习题	168
第 9 章 数字合不合理?	174
他们在说什么?	176
数字彼此之间是否相符?	177
数字可信吗?	178
数字是否好得不像真的?	179
算术对不对?	180
背后有什么该注意的吗?	183
网络寻奇	184
本章重点摘要	185
第 9 章 习题	186
第一部分 复习	192
第一部分 重点摘要	194
第一部分 复习习题	196
第一部分 报告作业	202
第二部分 整合数据	205
第 10 章 好的图及坏的图	206
数据表	208
饼状图及柱状图	210
留意象形图	213
随着时间变动的线图	214
注意刻度	216
怎样把图画好	217
本章重点摘要	220
第 10 章 习题	221



第11章 用图形呈现分布	232
直方图	234
解释直方图	237
茎叶图	241
本章重点摘要	244
第11章 习题	245
第12章 用数字描述分布	254
中位数和四分位数	256
五数综合及箱形图	259
统计学上的争议：贫富差距	262
平均数和标准差	264
选择数值描述	268
网络寻奇	270
本章重点摘要	271
第12章 习题	272
第13章 正态分布	282
密度曲线	285
密度曲线的中心和离度	286
正态分布	287
68-95-99.7 规则	290
标准计分	292
正态分布的百分位数*	294
网络寻奇	296
本章重点摘要	297
第13章 习题	298

*选读部分。





第 14 章 描述相关关系的方法： 散布图和相关系数	306
散布图	309
诠释散布图	310
相关系数	313
了解相关系数的意义	315
网络寻奇	318
本章重点摘要	319
第 14 章 习题	320
 第 15 章 描述相关关系： 回归、预测及因果关系	 330
回归直线	331
回归方程式	333
了解预测的意义	336
相关系数及回归	338
因果问题	340
统计学上的争议：枪械管制和犯罪	342
因果证据	345
网络寻奇	346
本章重点摘要	347
第 15 章 习题	348
 第 16 章 消费者物价指数和政府统计	 358
指数	360
固定市场总览物价指数	361
如何使用 CPI	363
了解 CPI 的意义	367
统计学上的争议：CPI 把通货膨胀夸大了吗？	369



政府统计的处境	370
社会统计的问题	372
网络寻奇	373
本章重点摘要	374
第 16 章 习题	375

第二部分 复习	382
第二部分 重点摘要	385
第二部分 复习习题	388
第二部分 报告作业	396

第三部分 机遇	399
---------	-----

第 17 章 考虑可能性	400
概率的概念	401
机遇的古代史	404
关于机遇结果的神话	405
个人概率	410
概率及风险	411
网络寻奇	413
本章重点摘要	414
第 17 章 习题	415

第 18 章 概率模型	420
概率模型	421
概率规则	422
抽样的概率模型	425
本章重点摘要	429
第 18 章 习题	430





第 19 章 模拟	436
概率从何而来?	437
模拟入门	438
更复杂的模拟	441
网络寻奇	445
本章重点摘要	446
第 19 章 习题	447
第 20 章 赌场的优势: 期望值	454
期望值	455
大数法则	459
深入探讨期望值	459
用模拟计算期望值	460
统计学上的争议: 合法赌博面面观	462
网络寻奇	463
本章重点摘要	464
第 20 章 习题	465
第三部分 复习	472
第三部分 重点摘要	474
第三部分 复习习题	476
第三部分 报告作业	481
第四部分 推论	483
第 21 章 什么是置信区间	484
估计	485
有信心的估计	487
了解置信区间	491



总体比例的置信区间*	494
本章重点摘要	497
第 21 章 习题	498
第 22 章 什么是显著性检验	504
统计检验的理论基础	505
假设及 P 值	508
统计显著性	512
计算 P 值*	513
网络寻奇	514
本章重点摘要	515
第 22 章 习题	516
第 23 章 统计推论的使用与滥用	522
聪明做推论	523
显著性检验面临的难处	525
置信区间的优点	529
统计学上的争议：应不应该禁止统计检验？	530
“5% 的显著水平”并非魔术指标	531
提防刻意寻找的显著性	531
网络寻奇	533
本章重点摘要	534
第 23 章 习题	535
第 24 章 双向表及卡方检验*	540
双向表	541
辛浦森悖论	543
双向表的推论	545
卡方检验	548
如何应用卡方检验	552
本章重点摘要	555



选读部分。





第 24 章 习题	556
第 25 章 有关总体平均数的推论*	562
样本平均数的抽样分布	563
总体平均数的置信区间	566
总体平均的检验	568
本章重点摘要	572
第 25 章 习题	573
第四部分 复习	578
第四部分 重点摘要	580
第四部分 复习习题	582
第四部分 报告作业	590
 注释与资料出处	 592
 部分习题解答	 614
 表 A 随机数字	 632
 表 B 正态分布的百分位数	 634

* 选读部分。

第一部分

产生数据

你和你的朋友不是典型人物。比如说你喜欢的音乐类型，可能就和我喜欢的不一样。当然我和我的朋友也一样不是典型的。如果要知道整个国家(或只是大学生)的状况，我们必须认清一个事实，就是整体状况也许并不接近我们自己，或我们周围的状况。所以我们需要数据。从零售店得到的资料显示，美国最畅销的音乐是蓝调(1999年卖了1.75亿张唱片)以及另类音乐(1.21亿张)。如果你喜欢重金属(3000万张)而我喜欢爵士乐(只有200万张)，我们可能对整个唱片消费群体的音乐品味毫无所悉。如果我们身处唱片业，或者只是对通俗文化感兴趣，就必须把我们自己的喜好摆在一边，而好好地来检视数据。

你可以到图书馆或者上网找数据(唱片销售量就是我上网找来的)。但是我们怎么知道数据可不可靠呢?好的数据也可以看成像毛衣或数码影音光碟机(DVD)之类的产品。草率产生的数据就像草率生产的毛衣或光碟机一样令人沮丧。你买毛衣前会检查一下，如果做得很差你不会买。数据也一样，如果“制作”得很糟你就不该用。这本书的第一个部分就会告诉你怎样分辨数据好不好。

第 1 章

数据从何而来?

光看不想，看不出所以然

你可能读了几个月报纸、看了几个月电视新闻，都没有遇到过任何代数公式，难怪你会觉得代数好像和实际生活无关。不过没有任何一天，你会完全没有接触到数据和统计研究。你听说上个月的失业率是 4.5%。报纸上报道说，年龄介于 18—29 岁之间的人，只有 21% 声称他们经常投票，而 65 岁以上的人则有 59% 这样说。还有一篇更长些的报道中说，低收入户儿童若有良好的日间照顾，则和其他同样背景的儿童比起来，数年之后的考试成绩会比较好，读大学的概率比较大，也有较好的工作。

这些数据是打哪儿来的？为什么我们应该相信？或者我们也许并不



该相信。也许是像约吉·贝拉(Yogi Berra)曾说过的：“你只要肯看，就可以观察到许多事。”但是你怎么看也不可能看出，年轻人的投票率只有 21%；或者良好的日间照顾，会使儿童在 15 年之后进入大学就读。好数据是人们智慧及努力的产物。坏数据的来源，则是懒惰、不了解甚至存心误导。每当有人丢个数字给你，你第一个该问的问题就是：“这数字是打哪儿来的？”

来说说数据：个体和变量

统计是数据的科学。我几乎要说它是“数据的艺术”了，因为除了要有好的判断外，甚至还要有好的品味，再加上好的数学，才能造就好的统计。好的判断中有一大部分，是在于决定你究竟要度量什么，如此产生的数据才能帮你了解你所关心的问题。我们得先介绍一些名词，它们是用来描述、组成数据的原始材料。

• 个体及变量

个体(individual)就是一组数据描绘的对象。个体也许是指人，但也可以是动物或其他东西。

变量(variable)是指一个个体的任意“特征”(characteristic)，同一个变量对于不同个体，可能有不同的值。

举例来说，以下是一门统计课程结束时，教授记分簿资料中的头几列：

姓名	主修	分数	等级
业樊尼	大众传播	397	B
巴顿	历史	323	C
布朗	文学	446	A
邱森	心理	405	B
柯堤兹	心理	461	A



这个例子中的个体，就是修课的学生。除了学生姓名之外，还有三个变量。第一个变量告诉我们学生主修什么。第二个变量是学生在整个课程的 500 分总分中，共得了几分。而第三个变量就是学生在这科所得的学期成绩。

统计是处理数字的，但是并非所有变量的值都是数值。在这位教授的记分簿资料中的三个变量，只有分数这个变量的值是数值。要用统计来处理其他非数值变量时，可以用计数(count)或百分比(percent)。比如说我们可以列出有多少百分比的学生得到 A，或者主修心理的学生所占比例。

选择变量时若判断不正确，可能导致花大笔钱和时间来取得数据，而这些数据却没什么用处。但到底什么才是正确的判断，可能并无定论。以下的例子告诉我们，决定要搜集何种资料时所会面对的挑战。

例 1 谁做资源回收？

谁不嫌麻烦做资源回收？调查人员花了许多时间和金钱，在加州某一城市的两个地区，把居民拿到屋外等着回收的东西过磅。我们姑且称这两区为上层区和中低区。这里的个体就是住户(household)，因为收垃圾和回收物资是为整户居民做的，而不是分别对每一个人做。所度量的变量就是放在路边的资源回收垃圾每周的重量。

上层区的住户，平均来说每周贡献的重量，都超过中低区的居民。我们是不是可以说，有钱人较认真地做资源回收呢？答案是否定的。有人注意到上层区的垃圾桶里，有很多很重的玻璃酒瓶。而在中低区，拿出来的很多都是很轻的塑料汽水瓶、啤酒罐及汽水瓶。所以光凭重量判断，对于谁比较用心做资源回收，所知有限。

观测研究

2

有的时候你只能观测。如果要知道黑猩猩在野地里的行为，你得观察。要研究老师和小朋友在教室中的互动，你也得观察。若观察者



知道自己要看的是什么,就会容易一些。黑猩猩专家可能对雌猩猩和雄猩猩的相互影响有兴趣,或者想知道是否猩猩群中的某几只拥有统治权,还有黑猩猩会不会猎食动物的肉。事实上大家一直以为黑猩猩是素食动物,直到简·古道尔(Jane Goodall)在坦桑尼亚(Tanzania)的冈贝国家公园(Gombe National Park)仔细观察它们之后才改变想法。现在已清楚地知道,肉是黑猩猩日常食物的一部分。

开始的时候,观察者可能不知道应该记录些什么。然而迟早会有些模式出现,我们就可以决定应该度量哪些变量。比如黑猩猩多久猎食一次?单独行动还是群体行动?多少只一起行动?只有雄性还是雌,雄都有?食物中肉类占了多少比重?有系统的观察,加上清楚定义所度量的变量,会比光是观察更有说服力。以下是一个规划完善(且很花钱)的观察研究案例。

例2 高压电线会使儿童得白血病吗?

电流会产生磁场,所以生活在有电的环境里,会使人暴露在磁场中。住在高压电线附近,会增加这种暴露的程度。实验室中的研究显示,强烈的磁场会干扰活细胞(living cell),但是因为住在高压电线附近,而接触到较弱的磁场,影响又如何呢?有些数据显示,似乎住在这些地方的儿童,会有较多的人患上属于血癌的白血病。

我们不能安排孩子去暴露在磁场下来做实验。而要比较暴露较多和较少磁场下的儿童罹患白血病的比例也有点困难,因为白血病很罕见,而儿童居住的位置除了磁场暴露程度不同之外,也可能有许多其他的差异。可行的方法是从已得白血病的儿童着手,把他们和未得白血病的儿童做比较。我们可以检视许多的可能原因,像食物、杀虫剂、饮水、磁场等等,看看有白血病和无白血病的儿童,在这些项目中,有哪些不同。在这些大规模的研究中,有一些显示出,似乎应该对磁场做进一步的研究。

结果有人花了5年的时间和500万美元,对磁场做了极为仔细的研究。研究者将638个得白血病的儿童和620个未得病的做比较,他们到这些儿童的家里,在儿童的卧房、其他房间及房子的前门处都测量了磁场的强度。不仅对儿童住家附近的高压电线资料做了记录,还对儿童的母亲在怀孕时的住处附近的高压电线也做了记录。结论是,除了巧合之外,并没有证据显示磁场和儿童白血病有相关关系。



“没有证据”显示磁场和儿童白血病有关，并不是证明暴露在磁场下完全没有风险。而只是说经过非常严谨的研究之后，并没有证据可以做出结论说，磁场有导致白血病风险的机会。有些评论持续批评，认为这个研究漏掉了一些重要的变量，或者参与研究的儿童不具代表性。不过不论怎么说，一个经过详尽规划的观测研究，肯定是要比随随便便，甚至有时情绪化的提出的几个癌症病例，要令人信服得多了。

• 观测研究

观测研究(observational study)是观察一些个体，并度量(measure)我们感兴趣的变量，但并不试图影响回应。观测研究的目的是描述一个团体或一种状况。

抽样调查

有句谚语说：“你不必吃完整头牛，才知道肉是老的。”这就是抽样的精髓：从检查一部分来得知全体。**抽样调查**(sample survey)是很重要的一种观测研究。他们只研究调查对象当中的一部分人，而选中这些人并不是因为对他们特别感兴趣，而是因为他们具代表性。以下是用来讨论抽样的词汇。

• 总体和样本

统计研究中的**总体**(population)，是指我们求取信息的对象全体。

样本(sample)是总体的一部分，我们从样本搜集信息，以便对整个总体做某些结论。

请注意，总体就是我们想研究的对象全体。如果我们想要得到关于美国所有大专学生的信息，那么所有美国大专学生就是我们的总体；即使选取样本时，也许受限制只能从一所大学里抽样，总体仍然不变。要想从样本中得出什么结论来，必须先知道该样本代表的总体是什么。比如说，选前民意调查到底问了哪些人的意见？是所有成年人？美国公民？已登记的选民？还是只问了民主党党员？样本只包括我们取



得信息的那些人。如果在调查当中,有些被选中的人联络不上,那么这些人就不包括在样本中。

总体和样本的区别,在统计里是很基本的观念。以下的例子会说明这个区别,并会介绍抽样的一些主要用途,同时也会指出我们对样本中个体度量的各种变量。

例3 民意调查

盖洛普(Gallup)及许多新闻机构常举办民意调查,探询人们对某些议题的意见。此处的变量,就是人们对公共政策相关问题的回答。这类民意调查整年都持续进行,但到了选举前才特别受注意。一个典型的民意调查,其总体及样本可能是以下状况:

总体: 18岁以上的美国居民,包括非公民,甚至非法移民。

样本: 从总体中选出且经过电话访谈的人,其人数在1 000—1 500之间。

例3 当前人口调查

美国政府的经济和社会资料,来自对全国的个人、住户或企业做大规模的抽样调查。美国最重要的政府抽样调查,是按月执行的“当前人口调查”(CPS, Current Population Survey)。CPS所记录的资料中,有许多资料和16岁以上人口是否就业有关。美国政府公布的每月失业率就是从CPS中得来的。CPS也记录许多其他的经济和社会资料。对于CPS来说:

总体: 超过一亿的全部美国住户。请注意,这里的个体是住户,而不是个人或家族。一个住户的组成分子,是所有住在同一个屋子中的人,而不论他们之间是何种关系。

样本: 每月所访谈的约50 000个住户。



你就是不懂

一项对新闻工作者和科学家所做的调查,发现两者之间在沟通时有隔阂。新闻工作者觉得科学家傲慢,而科学家认为新闻工作者无知。我们保持中立,不过调查中有一项有趣的结果:科学家中有 82% 同意,不管是医药或其他领域都会发生“由于媒体对统计的了解不够,以致无法解释新发现”。

例 5 电视收视率

“市场调查”是为了解消费者的喜好及产品的使用情形。市场调查中的一个著名例子是“尼尔森媒体研究”(Nielsen Media Research)做的电视收视率调查服务。尼尔森收视率影响广告商愿意花多少钱来买某节目的广告,以及该节目播不播得下去。对于尼尔森全国电视收视率来说:

总体: 所有一亿户有电视机的美国住户。

样本: 约 5 000 个住户,住户同意使用“个人收视记录器”(people meter)来记录该户中每个人收视的节日。所记录的变量包括住户中的人数与他们的年龄及性别,电视机开着的时段,及电视机开着时,是谁在看、看什么节目。

例 6 全面社会调查

社会科学研究经常使用抽样。设于芝加哥大学的“美国全国民意调查中心”(National Opinion Research Center)每两年做一次的“全面社会调查”(GSS, General Social Survey),是最重要的社会科学抽样调查。

其中所考虑的“变量”包括受访对象的个人及家庭背景、经验与习惯以及对某些主题的态度及意见。主题从堕胎到战争都有。

总体: 住在美国住户中的成年人(18 岁以上),不包括住在机构里的成年人:例如监狱里的犯人及住在大学宿舍里的人,也不包括无法以英语访谈的人。

样本: 约 3 000 个成人,访谈是面对面在受访者的住所进行。



大部分统计研究中用的样本是“广义”的样本。比如说，在例2中的638个白血病患者，是被视为所有白血病患者的代表。通常“抽样调查”这样尊贵的词，我们是保留给有计划的研究中，抽取样本来代表某个总体的时候才用。例2中的白血病患者，是专门治疗儿童癌症的治疗中心的病人。专家判断说，这些儿童可以代表所有的白血病患者，即使他们全都是一些高级医院的病人。抽样调查可不是靠判断来做的，抽样调查要从一个总体开始，然后从中抽取一个能代表总体的样本。在第2章到第4章里，就要讨论抽样调查的艺术及科学。

普查

抽样调查只看总体的一部分，为什么不看全部呢？普查就是想看全部。

• 普查

普查(census)是企图把整个总体纳入样本的抽样调查。

美国宪法规定每10年要做一次全国人口调查。要对这么大的总体做普查，既费钱又费时。即使联邦政府负担得起普查的费用，仍然还是利用样本，比如像CPS，来获取失业率及其他许多变量的及时资料。若政府真的要问全国每一个成人的工作状况，那么这个月的失业率恐怕要等到明年才会知道。即使在每10年一次的普查当中，仍然包括了一个抽样调查。他们从所有住户中抽了六分之一当样本，给其中每户寄了一份“繁式”普查问卷，上面问的问题比基本的普查问卷多得多。

所以从时间和金钱的角度来看，抽样比普查划算，而且抽样还有其他优点。假如你要测试鞭炮或保险丝的功能是否正常，测试过的产品就都毁了。还有，比起全面普查，较小的样本反而可能会得到较精确的结果。要职员去清点库存的所有50万个备份零件，不如要他去取一个样本好好清点。因为职员数烦的时候，就会愈数愈不准了。

美国普查局的经验提醒我们：普查只能“试图”把整个总体纳入样本。普查局估计，1990年的全美普查漏掉了1.8%的美国人。漏掉

普查是否过时了？

美国从1790年开始，每10年就普查一次，但是科技日新月异，因此全国普查很有可能找到替代品。丹麦没有普查，法国也准备取消普查。丹麦对全国居民登记，居民有身份证，而且只要搬家就得变更登记。法国将要用一个大型抽样调查取代普查，这个调查将在不同的区域轮流执行。

美国普查局也有类似的想法。美国社区调查(American Community Survey)已经上路，而且繁式普查问卷在2000年之后就要取消了。



的人估计包括黑人族裔的 4.4%，且大多住在内陆城市。

即使有政府的强大资源做后盾，普查也不是一定做得到的。那到底为什么要普查呢？政府需要每个街区(block)的详细资料，才能划分出人数相同的选区。美国普查的主要作用，就是提供这些地方资料。

实 验

取样本的目的，是要了解总体的真实情况，而且在搜集信息时要尽量不产生干扰。所有观测研究都遵循同一原则：“观测，但别干扰。”当简·古道尔刚开始在坦桑尼亚观察黑猩猩时，曾设立了一个食物补给站，让黑猩猩在那儿可以吃到香蕉。稍后她说这个做法错了，因为这样有可能会改变猩猩的行为。



“现在开始吃香蕉，那可爱的统计学家正看着我们呢。”

从另一方面来看，当我们从事实验(experiment)时，却是存心要改变行为。在做实验时，我们不是只观察个体或问他们问题，而是刻意加上某些处理(treatment)，以期能够观察其反应。实验可以帮助我们解答诸如以下问题：“阿司匹林能减低心脏病发作的风险吗？”以及“如果让大学生在看不到商标的情形下，品尝百事可乐和可口可乐，大部分学生会较喜欢百事可乐吗？”



• 实验

实验(experiment)时会刻意对某些个体加上某项处理(treatment)，以期能够观察其反应。实验的目的是要研究，是否该特定处理会使反应改变。

例7 帮助领福利金的母亲找工作

大部分领福利金的成人，是有幼儿的母亲。对福利金妈妈做的观测研究显示，大部分人有能力可以增加收入，脱离领福利金的行列。有些人会利用自愿参加的工作训练计划，来增进自己的工作技能。是不是应该要求所有体格健全的福利金妈妈，都参加工作训练和寻求工作的计划呢？观测研究没法子告诉我们这项政策的效果。就算我们所研究的妈妈，是从所有领福利金的妈妈中适当选出的样本，但这些会参加工作训练及找工作的人，和不会的人之间可能原本就有许多差别。举例来说，可以从资料得知，找工作的这些人受过较多教育，但也可能这些人有不同的价值观及动机，而这些特征是没法观测到的。

想要得知要求福利金妈妈参加工作训练，是否能帮助她们自立，就必须实际对这计划做试验。在妈妈们开始申请福利金时，从当中选两组相似的人，要求其中一组参加工作训练，但是对另一组不提供这项计划。这就是一项实验。若干年之后，比较两组人的收入以及工作记录，就可以看出，要求参加工作计划是否有预期的成效。

福利金的例子说明了实验比观测研究更占优势。原则上来说，实验可以为“因果关系”(cause and effect)提供良好的证据。如果我们适当地设计实验，就可以从两组很相似的福利金妈妈开始。每个妈妈，当然会和别的妈妈有年龄、教育程度、子女数及其他方面的差别。但是当我们检视两组当中所有人的年龄、教育程度、子女数时，这两个组是很接近的。实验过程当中，大家过的生活都不一样，但是两组之间只有一项系统性的差别(systematic difference)，就是一组参加了工作计划，另一组没有。大家都经过了同样的经济起飞或不景气，都一起经过了观念的改变等等。如果参与训练的那一组，在拥有



工作和赚钱方面，都远胜过没参加训练的那组，我们可以说，训练计划确实造成了这个令人愉快的结果。

实验可以提供好的证据，显示其某项处理的确造成某种反应，这是计划当中的重大概念之一。重大的概念就要附带一个重要的提醒：统计结论是对一群个体“平均来说”(on the average)的结论。但对于任何特定个体，可就没说什么。平均来说，参与训练计划的人的收入比没参加的人多。这就说明了计划的目标达成了，可是并不代表每个参与计划的人都会受益。重大概念也会引起大大的疑问：如果我们希望计划会增加收入，那只是让某些女性参加，却不让其他人参加，会不会有点不道德？在第5章和第6章中会说明怎样设计好的实验，而第7章讨论伦理议题。

网络寻奇

美国国家环境卫生科学研究院(National Institute of Environmental Health Sciences)对于暴露于高压电线，对健康造成的可能影响的报告，你可以在这个网址找到：www.niehs.nih.gov/em-frapid。

从例4到例7，其中所提到的抽样调查，都有自己的网站：

- 盖洛普调查(例3)：www.gallup.com/poll
- 当前人口调查(例4)：www.bls.gov/cps/home.htm
- 尼尔森媒体研究(例5)：www.nielsenmedia.com
- 全面社会调查(例6)：www.norc.uchicago.edu/gss/home-page.htm

我推荐盖洛普网站，他们有最新的调查结果，且对抽样调查方法也做了清楚的解说。



本章重点摘要

任何一个统计研究都会记录关于一些**个体**(人、动物或东西)的资料,也就是一个或多个**变量**的值。有些变量比如年龄和收入,值是数值的。其他有些变量,比如职业或性别,就不是数值的。要确定研究中的变量,度量的就是你想要的信息。

对于任何一项统计研究,你最需要知道的,是数据是如何产生的。**观测研究**在只观测不干扰的情形下搜集信息。抽样调查是观测研究当中很重要的一种。**抽样调查**是从某个特定**总体**中抽取**样本**,然后从样本中撷取关于整个总体的信息。**普查**试图取得总体中每个个体的信息。**实验**会对个体做某件事情,然后观察个体如何反应。实验的目的通常是要了解,某种处理是否确实会引起某种反应。



第1章 习题

1.1 每加仑英里数。以下是描述 2000 年型汽车耗油状况(MPG, 每加仑英里数)的数据集(data set)的一小部分:

厂牌及车型	车型	排挡种类	汽缸数	城市耗油(MPG)	高速路耗油(MPG)
BMW328CI	小型车	自排	6	18	27
BMW328CI	小型车	手排	6	20	29
别克 Regal	中型车	自排	6	20	29
雪佛兰 Blazer	运动型多功能车	自排	6	16	20

- 这个数据集里的个体是什么?
- 对每个个体、度量了哪些变量?这些变量中哪些是数值的?

1.2 运动员的薪水。以下是职业棒球大联盟在 1999 年赛季第一天的部分球员资料:

球员	所属队	守备位置	年龄	年薪
邓伍迪	马林鱼	外野手	24	222
欧苏纳	道奇	投手	26	1 050
裴提特	洋基	投手	26	5 950
索沙	小熊	外野手	30	9 000

- 这个数据集描述的个体是什么?
- 除了球员姓名外,这个数据集中还包括几个变量?哪些变量是数值的?
- 你觉得这里的数值变量的单位(unit)是什么?比如说,索沙年薪 9 000 是代表多少钱?

1.3 谁做资源回收?在例 1 当中我们谈到,在同一个城市的两个不同地区间,比较资源回收的成效时,用重量来度量并不是理想的方法。你建议用什么变量来代替重量?

1.4 在职业女性中抽样。一位社会学家想知道,成年职业女性对于



政府的托儿补助有什么意见。她从当地一个企业及专业女性俱乐部拿到 520 个会员的名单, 从这 520 人中, 随机(at random)抽出 100 人寄问卷给她们, 但只回收了 48 份问卷。此研究的总体是什么? 实际取得信息的样本是什么? 社会学家想要取得信息的女性中, 有多少百分比有回应?

1.5 给总统打分数 报纸上报道一项民意调查的结果说: “43% 的美国人总统的整体表现感到满意。”报道最后写着: “这份调查是根据电话访问 1 210 位成人所得, 访问对象遍布美国各地, 但不包括阿拉斯加和夏威夷。”这个调查中度量的变量是什么? 你觉得报纸感兴趣, 想要获取信息的总体是什么? 样本是什么?

1.6 政党倾向的性别差异 在美国政党倾向似乎有性别差异, 因为女性倾向民主党的机会比男性要大。一位政治学者访问了许多登记选民, 其中两种性别都有, 她问每个人, 上次国会选举时, 票投给了民主党或共和党。这项研究是不是实验? 你的理由是什么? 研究中度量了什么变量?

1.7 总体是什么? 针对下面几小题抽取样本的状况, 将总体尽可能明确的指出来。也就是要说明, 总体是由哪些个体组成, 而且明确规范, 怎样的个体才会属于总体。如果题目提供的信息不足, 则合理的描述总体即可。

- (a) 一项民意调查联络了 1 161 位成人, 并且问他们: “你认为哪个政党对于如何在 21 世纪领导国家比较有想法?”
- (b) 一位家具制造商购买大批的硬木材。供应商应该在运货之前, 先把木材弄干, 因为湿木材的尺寸和形状会改变。家具制造商从每批货抽取 5 块木头来检验湿度。如果其中任何一块木头的湿度超过 12%, 就把整批货退回。
- (c) 美国社区调查将联络 3 百万美国住户, 全美每一个县, 都有一些住户会被包括进去。普查局这项新的调查, 会问每个住户有关居住状况、经济情况及社会地位的问题。这个新的调查, 将会取代原有的“繁式”普查问卷。

1.8 总体是什么? 针对下面几小题抽取样本的状况, 把总体尽可能明确的指出来。也就是要说明, 总体是由哪些个体组成, 而且明确规



范, 怎样的个体才会属于总体。如果题目提供的信息不足, 则合理的描述总体即可。

- (a) 一位商学院的研究者想要知道, 是哪些因素影响小型企业的存活与成功。她从大城市的分类电话簿中, 选出 150 家餐饮业。
- (b) 地区电视台想要知道, 电视观众是比较想看该地区自己的大学篮球队的比赛, 还是同时段进行的 NBA 比赛。他们播放了 NBA 篮球赛, 结果接到了 89 个电话, 要求播放自己的篮球队的比赛。
- (c) 一家保险公司想要监测, 该公司对于汽车投保户申报出险时的处理程序, 是否合乎标准。每个月公司都从该月收到的车险出险申请中抽取样本出来, 检查处理过程是否迅速确实。

1.9 公有住宅。为了要了解贫困户住在公有住宅是否有助于家庭稳定, 研究者取得了去年芝加哥所有申请公有住宅的申请人名单, 其中有些申请人顺利获准, 有些却被主管当局否决了。研究者访问了获准者及被拒者, 并做了比较。这是一项实验、抽样调查还是不属于抽样调查的观测研究? 请说明你的答案。

1.10 高压电线和白血病。例 2 对于高压电线和白血病关系的研究中, 比较了两组个体, 并且度量了许多可能影响两组之间差异的变量。请详细说明, 为何这项研究不是实验。

1.11 治疗前列腺疾病。一项大型研究使用了加拿大全民医疗系统的纪录资料, 来比较两种治疗前列腺疾病的方法哪一种较有效。这两种方法一是传统的手术治疗, 另一是不需手术的新方法。纪录中有许多病人的资料, 这些病人的医师, 有些选择了手术, 有些选择新疗法。研究显示, 使用新疗法的病人, 在 8 年内死亡的概率较高。

(a) 说明为何这是观测研究而非实验。

(b) 大略描述一下, 比较这两种疗法的实验, 应有怎样的特征。

1.12 运动和心脏病发作的关系。经常运动究竟能不能减低心脏病发作的风险? 以下是研究这个问题的两种方法。

- 1. 一位研究者找到 2 000 位年过 40 岁的男士, 他们都经常运动, 也未曾发作过心脏病。她为每个人“配”了一位各方面条件接近, 但是没有固定运动习惯的人, 然后观察运动组和非运动组长达 5 年的时间。



2. 另一位研究者找了4 000位40岁以上的男性,他们都没有发作过心脏病,也都愿意参与这项研究。她让其中2 000人参加了一项有人监督的定时运动计划,另外2 000人依照原来的习惯不做改变。研究者观察两组人长达5年的时间。
- (a) 说明为什么第一项研究是观测研究,而第二项是实验。
- (b) 为什么对于规律运动是否减低心脏病风险的问题上,实验可以提供更多有用的信息?

1.13 健康状况和领导能力 有一项对于健康状况(physical fitness)和领导能力之间关系的研究,选择的研究对象,是自愿参加某项运动计划的中年主管。这些主管依身体检查的结果,分成健康状况较佳组与健康状况较差组。然后每个人都参加一项为测试领导能力所设计的心理测验,再比较两组的测验结果。

- (a) 这是不是实验?请说明理由。
- (b) 我们觉得与其用自愿参加健康计划的人,不如做抽样调查。这项调查的调查者,看起来是对什么总体感兴趣?他们用了哪些变量?

1.14 汤姆·克兰西(Tom Clancy)用字不同的写作手法,有时可以经由用字的长短来分辨。有位对这方面感兴趣的学生,想要研究汤姆·克兰西在小说中用字的长度。她把一本汤姆·克兰西的小说随意翻到一页,并记录下那一页头250个字的长度。此研究中的总体是什么?样本是什么?学生度量的变量是什么?

1.15 选择你的研究种类。要回答下列问题,用哪种方法最好?要用实验、抽样调查或非抽样调查的观测研究?请说明你的选择。

- (a) 人家对国内的整体现状是否满意?
- (b) 大学生学基础会计,何者效果比较好?课堂上课还是网上教学?
- (c) 你的老师在课堂上问了问题之后,平均来说,等学生回答要等多久?

1.16 设定研究目标 举例说明,有关大学生或其行为或其想法的问题,最适合用以下何种方法找答案。

- (a) 抽样调查
- (b) 非抽样调查的观测研究
- (c) 实验。

第 2 章

好样本和坏样本

镇报做的民意调查

* 译注：路易斯安那州的堂区，相当于其他州的县。

在路易斯安那州的瑞皮德斯堂区* (Rapides Parish, Louisiana)，只有一家公司有权提供救护车服务。当地的报纸《镇报》(Town Talk) 要求读者打电话回应，来表达他们是否赞成让这家公司垄断。这类打电话回应通常采用自动化处理：赞成就打某个号码，不赞成则打另外一个。电话公司通常对打电话的人收费。

《镇报》共接到 3 763 个电话，显示出对于救护车超乎寻常的关切。调查之后发现，有 638 个电话来自救护车公司的办公室或公司高级主管的家里，而且无疑的，一定还有更多是较低阶层的员工打的。该公司的一位副总裁说：“我们有员工很关心这个状况，他们为工作



稳定性及家庭担心，所以可能多打了几个电话。”另有消息来源说，员工被嘱咐“早些投票、多多投票”，就像早年芝加哥黑帮控制选举时所说的一样。

我们以后会谈到的，3 763 这样大的一个样本已经是很足够的了，但前提是，样本的取得过程须合乎规范。邀请大家打电话（一打再打，打了又打），可不是个适当的抽样设计。

我们很快就会学到，专业人士如何抽取样本，以及这些方法为什么比《镇报》的方法好。

怎样可取得坏样本

《镇报》应该已经受到教训，明白取到坏样本比取到好样本来得容易。该报的民意调查依赖自发性回应 (voluntary response)，他们是要大家自己打电话进来，而不是主动抽取样本。结果就是有偏的 (biased)，样本里面赞成救护车垄断的比例，因此被加重了许多。自发性回应样本吸引到的，是对讨论中的议题有强烈感受的人。这些人，例如救护车公司的员工，可能并不能很公平地代表一般大众的意见。

要取得坏样本，不是只有上面的这种方法。比方说，我每个星期卖几箱橘子给你的公司，你从每箱当中抽几个橘子检查，以评定橘子的品质。最容易的做法是从摆在每箱最上面的橘子中抽取，但这些橘子可能无法代表整箱的情况，因为摆在底下的橘子较易在运送过程中损伤。假如我不够老实，也许会把烂橘子摆在底下，上面摆些好橘子让你检查。如果你从上面抽样，所得结果会是“有偏的”：样本橘子的品质总是优于他们所应代表的整个总体。



• 有偏抽样法

如果统计问题的设计使得结果总是往某个方向偏，我们就称这个设计是**有偏的**(biased)。

从总体抽样时，如果选最容易取得的，叫做**方便抽样**(convenience sampling)。

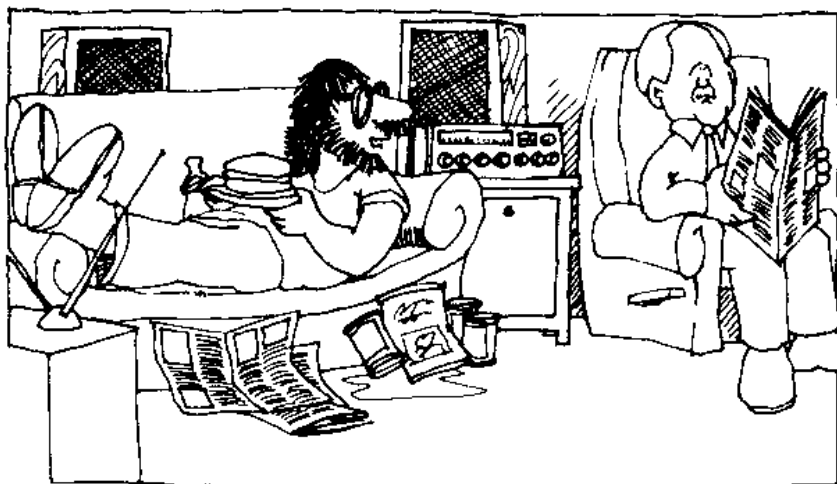
自发性回应样本(voluntary response sample)则是经由对某一诉求的回应而自然形成的。写信回应(write-in)或电话回应(call-in)意见调查都是自发性回应样本的例子。方便样本及自发性回应样本常常是有偏的。

例 1 购物中心访谈

只捏箱子里上层的橘子是方便抽样的一个例子，而在购物中心进行访谈是另一个例子。制造业者和广告代理商常常利用在购物中心的访谈，来搜集消费者的消费习惯及广告的效用等信息。在购物中心里取样本既快速又省钱，但在购物中心里访谈到的人并不能充分代表整个美国人口。比如说，这些人比较有钱，而且有很多青少年或退休人士。此外，访问者倾向于从顾客群中选择外表整洁、看起来不具威胁的人。购物中心的样本是有偏的，因为：某些群体的比重太重(较有钱的人、青少年及退休人士)，而有些群体的比重又太轻。这样一个方便样本的意见，可能和全美大众的意见有很大的出入。

例 2 写信回应意见调查

专栏作家蓝德丝(Ann Landers)有一次问她的读者：“如果可以重来一次，你还要孩子吗？”蓝德丝接到接近1万份答复，其中将近70%说：“不要！”难道说70%的父母都后悔有了孩子吗？当然不是。这是个自发性回应样本。通常对某个议题有强烈感觉的人，尤其是负面感觉的，比较会不嫌麻烦地去回应。蓝德丝的意见调查结果是有高度偏差的：她的样本中，宁愿不要孩子的父母百分比，远大于全体父母中宁愿不要孩子的百分比。



“嘿！老希，你昨天寄给蓝德兹的信里说了些什么？”

写信回应和电话回应的意见调查，几乎一定会导致有高度偏差的结果。事实上，只有 15% 的人曾经打电话去回复意见调查，而这些人可能也会打电话到广播电台的打电话回应节目去。对于整个人口来说，这些人并不构成具有代表性的样本。

简单随机样本

自发性回应样本，是由人们自行决定要不要回应；而方便样本则是由访谈者决定的。在这两种情形当中，都是由于人为因素而造成偏差。统计学家的补救方法，是利用不牵涉人为选择的“机遇” (chance) 来选取样本。用机遇选出的样本，既不会受取样者的偏好所影响，也不会有回应者的自行加入。用机遇选样本，是通过给每个个体有同样的中选机会，来消除偏差。不管有钱还是没钱，年轻还是年老，黑人还是白人，每个人被选进样本的机会都是一样的。

用机遇选样本最简单的方法是，把名字全部放进一顶帽子了（这就是总体），然后从当中抽出一部分（也就是样本）。这就是简单随机抽样 (simple random sampling) 的概念。

• 简单随机样本

大小为 n 的简单随机样本 (SRS, simple random sample) 是有 n 个个体的样本。其选取的方法，是使得总体中任一组 n 个个体，中选的的概率都相同。



SRS 不仅让每个个体有相同的中选机会(因此可消除选择偏差),也让每个可能的样本,有同样的中选机会。从帽子里抽名字就能做到这点:把 100 个名字分别写在同样大小的纸条上,放在帽子里混合均匀,这就是个总体;然后一张接一张,共抽出 10 张纸条,这就是一个 SRS,因为任何 10 张纸条和其他任何 10 张纸条,机会都一样。

每一个体或每一组可能的 n 个个体集合被抽中的概率相同,这个意义可以由帽子里抽名字中,得到清楚的理解。这就是 SRS 的涵义。当然如果我们想从全美 1 亿个住户中抽样,要从帽子里抽签就不大方便了,因此我们实际上是用电脑产生的随机数字(random digit)来选样本。如果不用电脑软件,你也可以利用随机数字表(table of random digits)来“人工”选取较小的样本。

• 随机数字

随机数字表(table of random digits)是一连串的 0、1、2、3、4、5、6、7、8、9 这些数字,且满足以下两个条件:

1. 表中任一位置的数字,其为 0—9 中任何一个数字的概率相同。
2. 不同位置的数字之间是独立的(independent)。也就是说,知道表中某一部分是些什么数字,不会提供你任何关于其他部分是些什么数字的信息。

书末的表 A 就是一个随机数字表。你可以想像成表 A 是这样来的:请一位助理(或叫电脑来做)把数字 0—9 放在一顶帽子当中混匀,随意抽出一个数字,记下来再放回,再混匀再抽,以此类推。助理先把混合及抽取的工作都做好了,所以我们要抽取随机数字时,就不必再重复进行这两项工作了。表 A 上一开头的数字是 19223950340575628713。为了让这个表容易读些,数字每 5 个排成一组,而且每列都有编号。这些“组”及“列”并没有特别意义,因为这个表只是一长串由随机选择而来的数字而已。现在我们举例说明,怎样用表 A 来选取 SRS。

利用随机数字表,比从帽子里抽签要快多了。就如例 3 中所表达的,选取 SRS 有两个步骤。



例3 如何选取 SRS

琼恩的小规模会计师事务所共有 30 家客户。为了要想办法增加客户满意度，琼恩想要选 5 家客户来访谈。为了避免偏差，琼恩决定选大小为 5 的 SRS。

步骤 1：编代码

对每家客户编一个数字代码，数字的位数尽量小，够用即可。30 个客户要用 2 位数，所以我们就用：

01, 02, 03, ..., 29, 30

要用 00—29 也可以，或者任何 30 个 2 位数都行。编好代码的客户清单如下页的表所示。

步骤 2：利用随机数字表

从表 A 中任一处开始，一次读 2 个数字。假如我们从编号 130 的列开始，该列数字为：

69051 64817 87174 09517 84534 06489
87201 97245

01 恩一水管工程	16 杰儿唱片
02 特色印刷	17 强生日用品
03 行动运动器材	18 凯瑟营造
04 安德森营造	19 刘氏中餐馆
05 贝利货运	20 神奇褐肤中心
06 飞船公司	21 无敌机械
07 班奈五金	22 艺术摄影
08 贝斯照相器材	23 河城图书
09 蓝图专门	24 河滨酒店
10 中央树木处理	25 乡村流行专卖店
11 古典花卉	26 卫星服务
12 电脑咨询	27 苏格兰洗衣店
13 达琳玩偶	28 污水处理
14 佛莱房地产	29 轮胎专门
15 赫南迪电子	30 冯恩录影带

这些随机数字真的是随机的吗？

才怪呢！表 A 中的随机数字是用电脑程序跑出来的，而电脑程序是完全遵命行事的。你只要对程序输入同样的东西，电脑就会产生同样的“随机”数字。当然啦，有些聪明的家伙把电脑程序设计得很高明，使得产生出来的数字很像是随机的。这些其实叫做“拟随机数”（pseudo-random numbers），而表 A 当中的数字就是属于这种。拟随机数用在统计上的随机化效果不错，但是一些更细腻的用法，可能就会被这些数字背后隐藏的不随机的形态给搞砸了。



其中头 10 个“2 位数字组”为

69 05 16 48 17 87 17 40 95 17

表 A 中任一个“2 位数字组”，其为 00、01、02……99 这 100 组数字中任一组的概率都相同。所以选 2 位数字组就等于随机选代码。这就是我们的目标。

琼思的代码只用了 01—30，所以在这以外的二位数，我们都不要。这样子得到的头 5 个在 01—30 之间的代码，就代表我们选出的样本。在列 130 的头 10 个代码中，有 5 个超过 30，这些我们扔掉不用，剩下的是 05、16、17、17 及 17。代码 05、16 及 17 的客户就被收进样本，我们就不理第 2 及第 3 个 17 了，因为 17 已经被选进样本中。照这样在列 130 继续搜寻下去（有必要可延续到 131 列），直到选好 5 个客户为止。

如此得到的样本，包括代码 05、16、17、20、19，分别代表贝利货运、杰儿唱片、强生日用品、神奇褐肤中心及刘氏中餐馆。

• 用两个步骤选取 SRS

步骤 1：编代码

对个体中每一个个体，指定一个数字代码。要确定每个代码都是同样的位数。

步骤 2：用表

利用随机数字来随机选代码。

指定代码可以用任何方便的方式来进行，比如英文姓名就可照字母顺序来选。只要所有代码的位数都相同，所有个体就有同样的中选机会。代码要尽量短：总体的组成分子如果不超过 10 个，则使用个位数就够了；如果在 11—100 个之间，要用 2 位数；101—1 000 个之间，就要用 3 个数字了，以此类推。我建议你养成习惯，编代码都从 1（或 01，或 001，视需要而定）开始。表 A 里的数字可以往任何方向读，横着读、直着读等等，因为表里的数字并没有顺序。我建议不妨横着读。

抽样调查是用电脑软件来抽取 SRS，但是所用的软件，也就是把例 3 中的步骤“自动化”而已。电脑不用去查随机数字表，因为它自己就可以随时产生出随机数字。



样本可靠吗?

《镇报》、蓝德丝以及一些小型访问都产生了样本。但是我们没法信任从这些样本得到的结果,因为这些样本产生的方式都会导致偏差。但对于从SRS得到的结果,我们的信心就大得多,因为样本的抽取根据机遇,而没有人有因素干扰,因此可以避免偏差产生。对于任何一个样本,要问的第一个问题就是:样本是不是随机抽取的?民意调查和其他一些抽样调查的执行者,如果是专业人士,就都是采用随机抽样的。

例4 盖洛普调查

一项有关抽烟习惯的盖洛普调查,问了以下问题:
“你在过去一周内,有没有抽过烟?”报纸报导调查结果为
“美国人只有23%抽烟”。我们应该先问,盖洛普的样本怎么来的。报道里面稍后提到“结果是根据电话访问1039位在全美国随机选取的18岁以上成人所得,访问进行时间是1999年9月23—26日”。

提供这样的资料,就已经开始赢得我们的信任了。盖洛普告诉了我们,它考虑的总体是什么(住在美国各地的18岁以上成人)。我们也知道样本大小是1039,而且最重要的是,样本是随机抽取的。当然还有别的问题需要讨论,我们也会很快会讨论到,但是至少已经听到了令人放心的“随机选取”几个字。

随机的高尔夫篇

随机抽取让大家机会均等,所以如果需要决定哪些幸运儿可以得到某些难得的机会,比如打一场高尔夫的时候,用随机抽取是一个公平的方法。许多高尔夫爱好者想要在位于苏格兰圣安德鲁斯的著名老球场(Old Course)打球。但只有少数人能预约得到。大部分人只能希望,在每天抽签决定谁能打时,能受幸运之神眷顾。在夏天旺季的时候,每6人中只有1人,可以得到花120美元打一场球的权利。



网络寻奇

互联网可以把一些自发性回应民意调查结果,送到你手边的电脑上。你可以进入 www.misterpoll.com, 并把自己变成数十个网上民意调查中任一个的样本之一。正如网站上说的,“这些民意调查没有一个是很科学的,但是的确代表所有参与者的集体意见。”

如果想知道如何用软件快速选取 SRS, 只要造访“研究用随机性发生器”(Research Randomizer)的网站:www.randomizer.org, 在首页上点击 Randomizer(随机性发生器)并填些资料即可。你甚至可以叫随机性发生器帮你把样本排序呢。



本章重点摘要

我们选取**样本**，以期得到有关**总体**的资讯。怎样可以选到较能代表总体的样本呢？**方便样本**和**自发性回应样本**都常有人使用，但是产生的结果常令人存疑。这些抽样方法通常是**有偏的**。也就是说，在选取样本时，会有系统的偏向于总体中的某一部分。

刻意利用机遇来产生数据，是统计当中的重大概念之一。随机样本用机遇来挑样本，因此可以避免人为选择的偏差。随机样本中最基本的一种是**简单随机样本(SRS)**，它选取的方式会使得所有同样大小的样本，都有同样的机会中选。要用人工选取 SRS 的话，可以用和本书末的表 A 一样的**随机数字表**。



第2章 习题

2.1 写信给美国国会。假设你是美国某国会议员的幕僚，这位议员正在考虑一项法案，该法案会对老人疗养院的服务，提供政府资助的保险。你的报告指出，一共收到 1 128 封针对此法案的来信，其中 871 封反对此项法案。国会议员说：“真没想到我的选区当中，大部分人都反对这个法案。我还以为会有很多人赞成。”你相信大部分的选民都反对这个法案吗？你会怎么向国会议员解释这件事牵涉到的统计问题？

2.2 即时意见调查。Harris/Excite 网站有即时意见调查，你可以查询网址：poll.excite.com/poll/home.jsp?cat_id=1。问题会出现在屏幕上，你只要点选“同意”、“不同意”或“不知道”即可。在 2000 年 1 月 25 日那天，问题是：“女运动员和男运动员是否应该同工同酬？”该调查接到的回答当中，有 13 147 个(44%)表示同意，15 182 个(50%)表示不同意，剩下的 1 448 个说不知道。

(a) 此调查的样本多人？

(b) 这个样本比一般抽样调查的标准样本大了许多。即使如此，我们还是不认为，这个结果对任何清楚定义的总体提供了有用的信息。为什么？

(c) 上网的人，目前仍然是男性比女性多，这件事对调查结果会有何影响？

2.3 蓝德丝女士取的样本。替读者提供建议的专栏作家蓝德丝，有一次问她的女性读者，是不是只要男性用柔情对待她们，即使没有性行为，也一样觉得满足。有超过 9 万名女性回应，其中 72% 回答“是。”在许多人的信中，描述了男人对她们的无情对待，说明为什么这个样本必定是有偏的。你觉得应该是往哪个方向偏？也就是说，72% 这个数字，和所有成年女性中真正的赞成比例相较，是较高还是较低？

2.4 我们不喜欢单行道。公路策划单位决定要把印第安纳州西拉斐特的一条主要道路变成单行道。《拉斐特日报》(*The Lafayette Jour-*



nal and Courier)进行了一天的民意调查,邀请读者打某部电话,表达他们的意见。第二天该报有如下的报道:

本报读者一面倒地认为,在西拉斐特的乡村区,应该有双向的交通,不该设单行道。周三打电话到本报表达意见专线的读者,大约8个人里面有7个,对5月份开始设立的单行道有所抱怨。在收到的98件意见中,没有反对单行道的只有14件

- (a) 你认为该报想要寻求什么样的总体的信息?
- (b) 这个总体中赞成单行道的比例,是否几乎一定大于、还是一定小于样本比例 $14/98$? 为什么?

2.5 自己设计坏样本 你读的大学想要搜集同学们对于在校内停车的意见,但不可能实际去问每一个学生的意见。

- (a) 举出一个坏抽样方法,该方法要依赖自发性回应
- (b) 另举一个坏抽样方法,这同不用自发性回应

2.6 打电话回应意见调查 联合国的总部,是否应该继续设在美国?有个电视节目对此议题邀请观众打电话回应表示意见,一共接到186 000个电话,其中67%说“不应该”。从全国抽的一个500个成人的随机样本中,却有72%的人对同样的问题回答“应该”。试试看怎样可以向一个不懂统计的人解释,要知道所有美国人是怎么想的,这500个随机选择的人的意见,反倒比186 000个自动打电话的人(次)的意见要来得可靠?

2.7 选取一个SRS 有家公司希望知道,公司的经理当中属于少数族裔的人,对于公司评估经理表现的系统,态度如何。底下是该公司所有属于少数族裔的经理名单。利用表A的列139选出6位,来仔细了解他们对表现评估系统的看法。

艾嘎瓦	狄涯德	黄	普瑞
艾方西卡	佛莱明	金	瑞查兹
拜克斯特	方西卡	廖	罗缀格
包曼	盖茨	墨宁	圣地亚哥
布朗	葛尔	努内慈	沈
柯帖兹	高梅兹	皮特斯	斐格
克罗丝	贺南德兹	皮里耶哥	渡边



2.8 选取一个 SRS。你修的“古代乌加里特宗教”这门课教得很差，你们班决定要向院长反映情况，大家同意要在班上随机选 4 个人去见院长。班上学生名单如下。用随机数字表来选一个大小为 4 的 SRS，从列 145 开始。

安德森	古提瑞兹	派乃克
艾斯平	葛温	波瑞里
贝尼特兹	哈特	饶欧
巴克	韩德森	瑞德
布莱曼	休斯	罗勃森
卡斯帝洛	江森	罗缀格
狄克森	坎卜桑	席格
爱德华	梁	汤普金斯
佛南地	蒙拓亚	范德葛瑞夫
辜普塔	欧兹	王

2.9 选举日样本。你想要在一个城市的 440 个选区中，选出一个包含 25 个选区的 SRS，将在选举日时严密监督选举过程的公正性。

- (a) 详细说明你要怎样给 440 个选区编代码。一个代码要用几位数？你用的位数，最多可以帮多少选区编代码？
- (b) 用表 A 来选出这个 SRS，并列出你选中的选区代码。你从表 A 的列 117 开始。

2.10 这是不是个 SRS？某大学有 1 000 位男教师及 500 位女教师。为了调查教师的意见，从 1 000 位男教师中随机抽取了 100 位，另外又从 500 位女教师中，随机抽取了 50 位。我们的样本就包括这 150 位抽出来的教师。

- (a) 说明为什么这样的抽样方法，让每位教师被抽到的机会是一样的。
- (b) 然而这可不是一个 SRS。为什么呢？

2.11 学生赚多少钱？某所大学里负责提供学生财政补贴的部门想要知道，学生在暑期工作里大约能赚到多少钱。这个信息会用来作为决定补助金额的参考。总体包括 3 478 位就读至少一年，而尚未毕业的学生。学校会从一个按字母顺序排序的名单中抽一个 100 人的 SRS，



然后寄问卷给他们填答。

- (a) 说明你会怎样给学生编代码,以便选出 SRS。
- (b) 用表 A,列 105,来选出样本中头 5 位学生。

2.12 公寓生活。你计划对某大学城里的公寓居民,做一份居住情况的报告。你决定要随机选出 3 个公寓社区,然后对其中的居民做深入的访问。利用表 A,由列 117 开始,从以下社区当中选出一个大小为 3 的 SRS。

橡树园	乡景	梅费尔庄
湾区	乡村别墅	诺伯山
美丽花园	峰景	潘柏丽宅
灌木丛	德林	派柏密
布兰登镇	费林登	费瑟伦
欧石南根	费尔威丘	富邑
褐石屋	法勒	塞格摩脊
柏百丽	富兰克林公园	塞勒姆楼
康桥	乔治城	乡园
昌西村	绿南	沃特福德宅
乡绅	乐屋	威廉斯堡

2.13 随机数字有些什么性质?对于随机数字表,以下这些叙述哪些正确?哪些不正确?要说明理由。

- (a) 每列 40 个数字里,正好有 4 个 0。
- (b) 每一对数字,都有 1% 的机会是 00。
- (c) 表里面不可能出现像 0000 这样 4 个连续的 0,因为这个模式太不随机了。

2.14 你来决定!《今日美国》(USA Today)在报上登广告:电视节目上曾说:

对手枪的管制是否应该更严格?你可以参加今晚的特别节目——打电话回应民意调查来影响决定。如果赞成,打 1-900-720-6181;反对则打 1-900-720-6182。头一分钟的收费是 50 美分。



说明为什么这个民意调查的结果几乎一定是有偏的。

2.15 再多些随机。大部分的抽样调查都是随机抽取住宅电话来打，但并不是一定对接电话的人进行访问，他们可能问有哪些成人在该住宅中，然后随机选取一位为样本。这样会有什么好处？

2.16 给警察打分。美国迈阿密警察局想知道，当地的黑人居民对警察的表现有何观感。某位社会学家准备了一些有关警察的问题。警察局从主要是黑人居住的区域，抽出一个包含 300 个地址的 SRS，再派出一位穿制服的黑人警官，到每个抽出的地址，去访问一位成人。

(a) 总体和样本各是什么？

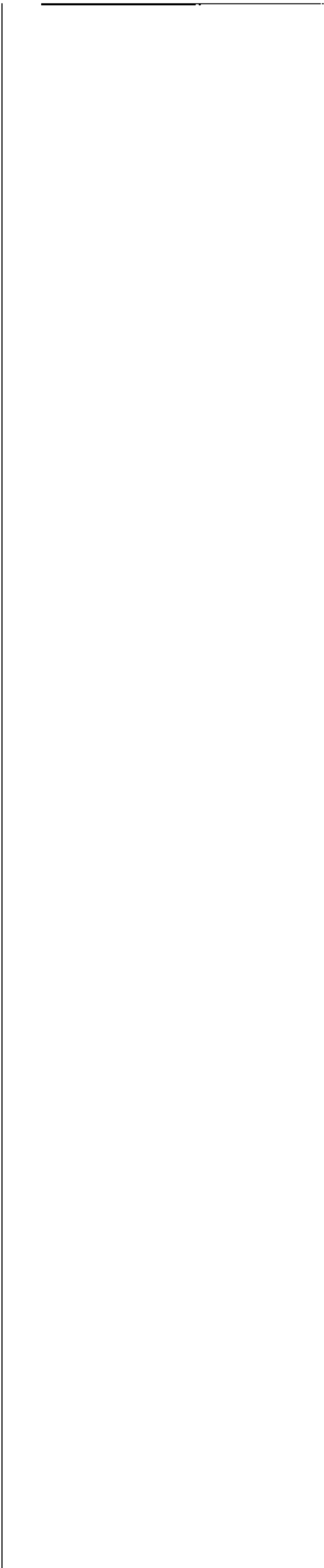
(b) 为什么虽然样本是 SRS，结果多半还是有偏的？

2.17 应该用随机选取吗？随机选取在多人竞争一样好东西时，是公平的决定方法，因为每个人赢的机会都一样。但是随机选择并非永远都是好主意，有时我们并不想对所有的人一视同仁，因为有的人或许比别人更有资格。在下面这些状况当中，你支持随机选取吗？请说出理由。

(a) 篮球比赛的场地有 4 000 个座位，但是有 7 000 个学生想要票。是不是应该在 7 000 人中随机选取 4 000 人？

(b) 等着换肝脏的病人，人数远超过能用来移植的肝脏数目，我们在决定把肝脏移植给谁时，应该完全用随机方式吗？

(c) 越战期间，是由抽签的随机方式，来决定年轻男子谁要去打仗。请问要决定谁得去越南，谁可以留在家，这是不是最好的方式？



第 3 章

样本告诉我们什么？

你玩乐透吗？

你知道乐透彩券在美国很受欢迎，不过到底有多受欢迎呢？盖洛普的报告中说：“乐透可能累积出高额奖金，而且彩券在你附近的店里就买得到，一张又只花 1 美元。对许多美国人来说，买张彩券已变成例行公事，尽管中奖概率微乎其微。最近一项以赌博为主题的盖洛普社会调查指出，过去 12 个月当中，有 57% 的美国人曾经购买过乐透彩券，这使得乐透成了当今赌博大众的最爱。”

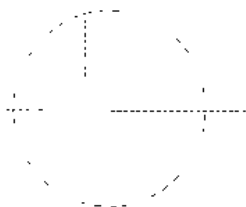
报告再读下去，我们还了解盖洛普是和 1 523 位随机选出的成人谈过之后，才得到这些结论的。我们很高兴盖洛普做了随机选择，要是我们走到排队买彩券的队伍面前去访问，恐怕就不会得到无偏的



(unbiased)结论了。但是慢着，人口普查局说美国大约有 2 亿成年人。即便它是一个随机样本，但光靠 1 523 个人的资料，怎么能告诉我们 2 亿人的习惯为何呢？

样本绝对没有办法告诉我们有关总体的准确信息。所以盖洛普调查都会提供我们一个“误差界限”(margin of error)。盖洛普会这样说：“对于用这样大小的样本所得到的结果来说，我们可以有 95% 的信心，由于抽样或其他随机因素所导致的误差，应在正负 3 个百分点之间。”

在新闻报道中，以上的叙述就简化为“此次民意调查的误差界限是正负 3 个百分点。”不管是哪种说法，到底是什么意思，好像都不是很清楚。想要完全了解的话，请继续读下去。



从样本到总体

盖洛普调查出“57% 美国人，在过去 12 个月当中，曾购买乐透彩券”，这是对于 2 亿成年人这个总体所做的声明。但是盖洛普并不知道整个总体的确实情况。他们的调查是访问了 1 523 位美国成人，并且得知其中有 57% 的人说，在过去一年当中曾经购买彩券。因为 1 523 位成人的这个样本是随机抽取的，如果认为样本相当可以代表总体，应该是合理的假设。所以盖洛普把“样本”中 57% 买了彩券的这个“事实”，转换成“所有成人”中约有 57% 买了彩券的这个“估计值”(estimate)。这是统计里面的一个基本动作：用样本的事实，当做总体真实信息的估计。要讨论这个主题之间，必须先分清楚哪个数字描述样本，哪个数字描述总体。以下是我们使用的词汇。



• 参数及统计量

参数(parameter)是描述总体的数字。参数是一个固定数字,但我们实际上无法知道参数的值。

统计量(statistic)是描述样本的数字。一旦取了样本,统计量的值就知道了,但是换个不同的样本,统计量的值就可能改变。我们常用统计量来估计未知的参数。

所以,参数之于总体,相当于统计量之于样本。想要估计未知的参数吗?只要从总体选一个样本,要用样本的统计量当做估计值就成了。盖洛普就是这么做的。

例 1 你玩乐透吗

所有的成年美国人在过去一年当中,买过乐透彩券的比例是一个参数,这个参数描述的是包含 2 亿成年人的总体。我们把这个比例用 p 表示,因为 p 是 proportion (比例)的第一个字母。但我们可不知道 p 的值。

为了估计 p ,盖洛普抽取了一个 1 523 人的样本。样本当中买了乐透彩券的比例就是一个统计量,我们称它为 \hat{p} (念成 p-hat)。结果 1 523 人的样本中,有 868 人买了彩券,所以对这个样本来说:

$$\hat{p} = \frac{868}{1\,523} = 0.57 \text{ (即 57\%)}$$

因为每个成人都有同样的机会被选进样本,所以如果用统计量 $\hat{p} = 0.57$ 当做未知参数 p 的估计值,似乎很合理。样本中有 57% 买了彩券是个事实,我们知道,因为我们问过样本中的每个人。我们并不知道所有成年人当中买彩券的比例,但是我们估计大概有 57% 的人买过。

抽样变异

如果盖洛普重新抽一个 1 523 个人的随机样本,这个样本会包含不一样的人。而且几乎也可以肯定,不会恰好有 868 人买过彩券。



也就是说, 统计量 \hat{p} 的值, 会随着样本的改变而改变, 因此可能会发生这样的情况: 一个随机样本说有 57% 的美国成人最近买过彩券, 而另一个随机样本却说只有 37% 的人买过。随机样本通过选样本的方法来消除偏差, 但是由于随机选取的结果会有变异, 所以得到的结果还是可能很不准。如果从同一总体重复取样, 但所得结果的变异太大的话, 我们就对任一个样本的结果都不敢信任了。

幸好随机样本的第二个大优点可以解救我们。它的第一大优点是, 随机选择可以消除“偏心”; 也就是说随机抽样把偏差给消灭了。第二大优点是, 如果我们从同一个总体, 重复抽取许多大小一样的随机样本, 所有样本的变异状况就会遵循某种可预测的形态(pattern)。从这个可预测的形态可以得知, 由较大样本所得结果的变异, 会小于小样本结果的变异。

例 2 很多样本

统计的另一个重要概念是这样的: 要知道一个样本的结果有多可靠, 就得先问问, 如果我们从同一个总体取很多样本, 会发生什么事情。我们来试试看。假设事实上(当然盖洛普并不知道)所有美国成人当中, 有 60% 在过去 12 个月里买过彩券。也就是说总体的真实比例是 $p = 0.6$ 。如果盖洛普用大小为 100 的 SRS 的样本比例 \hat{p} 来估计不知道的总体比例 p , 会发生什么情况?

图 3.1 就在表达抽很多样本, 并计算每一个样本的 \hat{p} 的过程。在第一个样本中, 100 个美国成年人里面有 56 人买了彩券, 所以 $\hat{p} = 56/100 = 0.56$, 下一个样本里只有 46 个人有买, 所以对这个样本来说, $\hat{p} = 0.46$ 。全部共抽取 1 000 个样本, 把这 1 000 个 \hat{p} 的值以“直方图”(histogram)表示, 就会得到图 3.1 右边的这个图: 图中 x 轴代表不同的 \hat{p} 值, 长方形的高度代表 1 000 个样本当中, 有多少个样本的 \hat{p} 值落在该长方形底部范围之内。

当然盖洛普共访问了 1 523 个人, 而不是只有 100 人。图 3.2 显示的, 就是从真实比例为 $p = 0.6$ 的总体所抽出, 大小为 1 523 的 1 000 个 SRS 所得到的结果。在图 3.1 和图 3.2 当中的两个直方图中, x 轴(横轴)的刻度一样, y 轴(纵轴)的刻度也一样。因此比较一下两个图形就可以看出来, 当样本大小从 100 增加到 1 523 的时候, 会发生什么情况。



仔细看看图 3.1 和 3.2。我们从总体开始，先抽出许多样本，然后从这些样本得到许多 \hat{p} 值。把这些 \hat{p} 值整合起来，就画出了直方图。现在来研究一下这两个直方图。

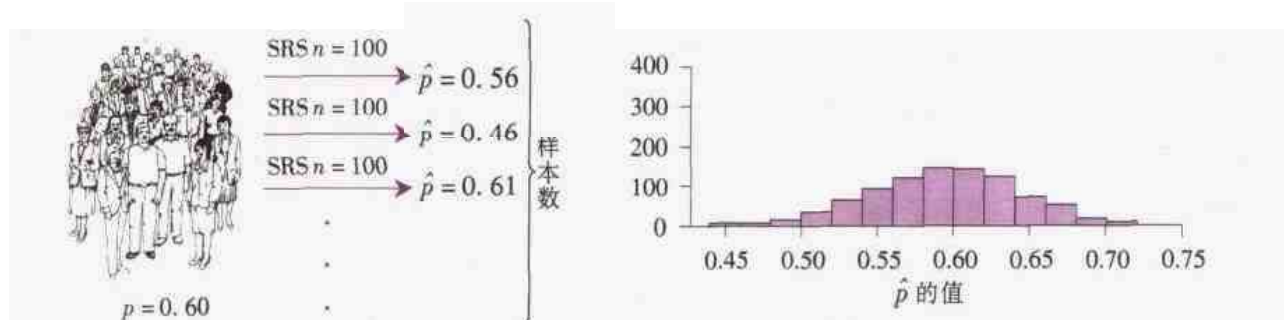


图 3.1 许多 SRS 的结果放在一起，会显现出某种有规则的形态。这里画的是从同一总体抽出的 1000 个大小为 100 的 SRS。总体比例为 $p = 0.60$ 。样本比例会随着样本而变，但是所有值的中心点，会落在总体的真实比例上

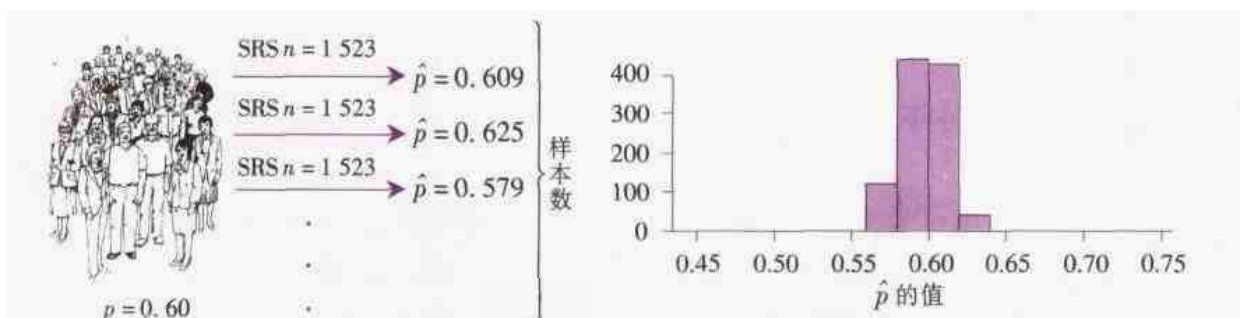


图 3.2 如同图 3.1 一样画直方图，只是样本大小改为 1523，仍然是取 1000 个 SRS。此处得到的 1000 个 \hat{p} 值，和图 3.1 的小样本结果比起来，散布范围窄得多，也就是较集中

- 不论样本的大小是 100 还是 1523，样本比例 \hat{p} 的值都会随不同的样本而变，但是这些 \hat{p} 值都以 0.6 为中心点。而前面提到过，0.6 是我们总体的真实比例。有些样本的 \hat{p} 值比 0.6 小，有些比 0.6 大，但是并不会有一部分都比较小，或大部分都比较大的倾向。也就是说，用 \hat{p} 值当作 p 的估计量 (estimator)，并没有偏差 (bias)。这点，不管样本是大是小都是如此。
- 大小为 100 的众多样本所算出的 \hat{p} 值，会比从大小为 1523 的众多样本所得到的 \hat{p} 值，要分散得多了。事实上，大小为 1523 的 1000 个样本当中，有 95% 的 \hat{p} 值在 0.576—0.624 之间。也就是与总体真实比例 0.6 差距在正负 0.024 的范围内。而大小为 100 的 1000 个样本，中间 95% 的值却分散在 0.50—0.69 之间，与真实比例约



有正负 0.1 的差距,差不多是刚才较大样本得到范围的 4 倍。所以大样本的**变异性**(variability)比小样本要小。

结论就是,我们可以指望一个大小为 1 523 的样本,其估计值 \hat{p} 几乎总会很靠近总体的真实比例。图 3.2 虽只针对一个特定的总体比例(即 0.6)说明这事实,但这对于任何总体而言都是正确的。而大小为 100 的样本,在真实比例是 60% 的时候,有可能得出 \hat{p} 为 50% 或 70% 的估计值。

想想图 3.1 和图 3.2 的意思,可以帮助我们重新整理一下,当我们用一个诸如 \hat{p} 的统计量,去估计诸如 p 这样的参数时,所谓的“偏差”是什么意思。同时也提醒了我们,变异性的重要程度不亚于偏差。

• 估计时的两种误差

偏差(bias)是当我们取很多样本时,统计量一直朝同一个方向偏离总体的参数值。

变异性(variability)描述的是,当我们取很多样本时,统计量的值会离散到什么程度。变异性大,就代表不同样本的结果可能差别很大。一个好的抽样方法,应该要有小偏差以及小变异性。

我们可以把总体参数的真正值想成是靶上的靶心,而把样本统计量想成是对着靶心发射的箭。偏差和变异性可以拿来形容弓箭手对着靶子射了许多箭之后的状况。

偏差的意思是我们的瞄准显然有问题,射出的箭都往同一个方向偏离靶心;样本值没有以总体值为中心点。

高变异性的意思是箭着点在靶子上分散得很广;重复抽样所得结果并不接近,彼此间差异很大。图 3.3 显示射靶的结果,说明这两种误差。

有没有注意到,即使是低变异性(箭孔都很接近),也可能有高偏差(箭孔都朝同一个方向偏离靶心);反过来说,即使偏差很小(箭孔呈现以靶心为中心点的散布),也可能伴随着高变异性(箭孔散布广)。好的抽样方法要像神箭手一样,必须同时具备低偏差及低变异性。

要达到这个目标,我们会这样做:

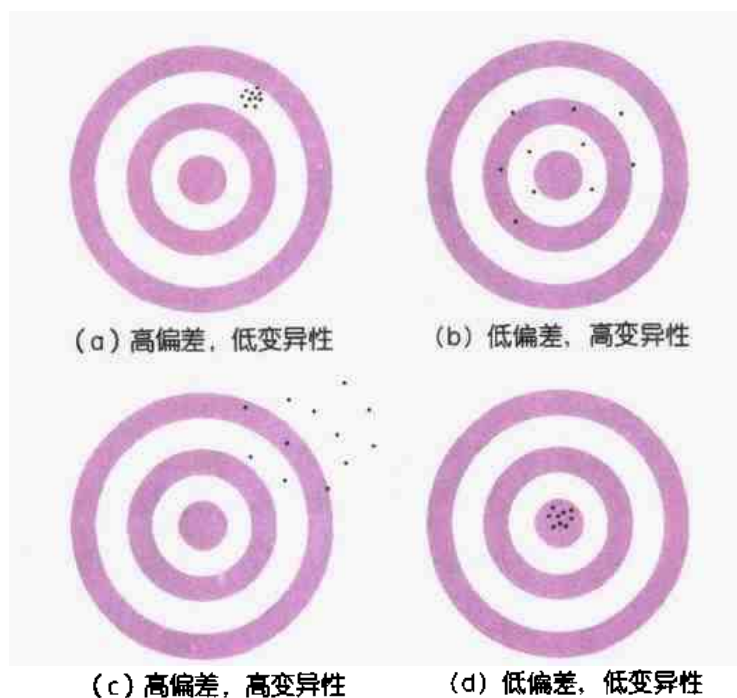


图 3.3 对着箭靶射箭时的偏差及变异性。偏差代表弓箭手老往同一个方向偏。变异性是指箭着点的分散情况

• 如何处理偏差及变异性

减低偏差：利用随机抽样即可。若先将整个总体列出来，再从中抽取简单随机样本，就会得到无偏估计值(unbiased estimate)，也就是说，以 SRS 得到的统计量来估计总体参数，既不会老是高估，也不会老是低估。

减低 SRS 的变异性：用大一点的样本。只要样本取得足够大，变异性要多小都可以做得到。

实际抽样的时候，盖洛普只取一个样本而已。我们不知道从这个样本得到的估计值，离真正值有多接近，因为我们根本不知道总体的真正值是多少。但是“只要是从大的随机样本算出的估计值，几乎一定会靠近真正值”。检视一下由许多样本结果构成的形态，使我们可以对一个样本的结果有信心。

误差界限

抽样调查报告中所宣告的“误差界限”，其实是把像图 3.1 及图



3.2 中所看到的抽样变异性,转换成一种我们对调查结果有多少信心的叙述来表达。我们先从在新闻中常听到的语言开始。

• 误差界限是什么意思

“误差界限(margin of error)是正负3个百分点”是以下叙述的缩写:

如果我们用和抽这个样本同样的方法,去抽许许多多样本,则这些样本中有95%,其所得的结果会在总体真正值的正负3个百分点之内。

让我们一步一步来看。通常一个随机样本的结果,不会刚好估计出总体的真正值。我们必须用一个误差界限,来表达我们的估计值距离真正值有多远。但是我们又不能百分之百确定,估计值和真正值的差距必定小于误差界限。所有样本当中有95%,距离真正值的确有这么近,但是另外的5%,距真正值的差距就超过误差界限了。我们并不知道总体的真正值是多少,所以我们也无法得知,到底我们的样本是属于那95%“中了”的样本,还是5%“没中”的样本。因此我们是说我们有95%的信心,真正值会在误差界限内。

例3 了解新闻内容

电视新闻播报员报道:“最新发布的盖洛普调查报告说,美国的成年人当中有57%,在过去12个月当中买过乐透彩券。此次调查的误差界限是3个百分点。”把57%加或减3个百分点,就得到54%—60%这个范围。大部分人以为,盖洛普宣称的是,整个总体的真正值,就落在这个范围里。

而盖洛普实际上说的是:“对于这样大小的样本所得的结果,我们可以说有95%的信心,由抽样或其他随机因素所造成的误差,应该是在正负3个百分点之内。也就是说,盖洛普告诉我们,误差界限只通用于95%的样本。“95%的信心”就是这种意思的精简说法。新闻报道中把“95%的信心”这句话给漏掉了。



确实计算出误差界限是统计学家要做的事。但是你可以用一个简单的公式，找出民意调查的误差界限大概有多大。

• 误差界限速算法

假设我们是在用大小为 n 的一个简单随机样本的样本比例 \hat{p} ，来估计未知的总体比例 p 。对应 95% 信心的误差界限，大致等于 $1/\sqrt{n}$ 。

例 4 误差界限是多少？

例 1 中的盖洛普调查访问了 1 523 人。对应 95% 信心的误差界限应该大约 是：

$$\frac{1}{\sqrt{1\,523}} = \frac{1}{39.03} = 0.026 \text{ (即 } 2.6\%)$$

盖洛普实际宣布的误差界限是 3%。我们的结果和盖洛普宣布的有一点差距是由于两个原因。首先，民意调查为了让新闻报道简单些，常常会宣布经过四舍五入到整数百分点的误差界限。第二，我们的速算公式适用于 SRS。在下一章里我们会看到，大部分全国性的样本比 SRS 复杂，其抽样方式通常会使得误差界限稍稍增大。不管怎样，我们的速算公式算出来的，已经相当接近现实了。

我们的速算法还透露出关于误差界限的重要信息。因为样本大小 n 是出现在公式的分母当中，所以较大的样本就有较小的误差界限。这个我们原本就知道，然而因为公式中用的是样本大小的平方根，所以若希望把误差界限减半，我们就得用一个 4 倍大的样本。

**例 5**

在例 2 当中, 我们把从同一总体所抽出的许多大小为 100 的 SRS, 和大小为 1 523 的 SRS 的结果做了比较, 我们发现中间 95% 的样本结果的分散状况, 小样本的误差界限约为大样本的 4 倍。

我们的速算公式估计出大小为 1 523 的 SRS 的误差界限, 差不多是 2.6%。而大小为 100 的 SRS 的误差界限, 大约是

$$\frac{1}{\sqrt{100}} = \frac{1}{10} = 0.1 (\text{即 } 10\%)$$

因为 1 523 很接近 16 乘以 100, 而 16 的平方根是 4, 所以 100 人的样本的误差界限, 差不多是 1 523 人的样本的 4 倍。

置信叙述

以下是盖洛普对于乐透彩券购买情况所做结论的精简版: “调查发现 57% 的美国成年人在过去 12 个月中曾购买彩券。我们有 95% 的信心, 所有美国成年人的真正购买比例, 会在这个样本结果的正负 3 个百分点范围内。”再来是超级精简版: “我们有 95% 的信心, 所有成年人当中, 有 54%—60% 曾在过去 12 个月里买过彩券。”这些都是置信叙述(confidence statement)。

■ 置信叙述

置信叙述(confidence statement)包含两个部分: **误差界限**(margin of error)及**置信水平**(level of confidence)。误差界限告诉我们, 样本统计量离总体参数多远。置信水平告诉我们, 所有可能样本中有多少百分比满足这样的误差界限。



如何拒绝电话推销

做抽样调查的人痛恨电话推销这档子事。我们都接过很多不想听的推销商品的电话，结果很多人在还没搞清楚对方不是在卖塑料墙板，而是在做抽样调查之前，就已经先挂了电话。在这儿教你一个分辨的诀窍：抽样调查和电话推销员都会随机选择电话号码，但是电话推销员会使用自动拨号系统打许许多多电话，在你接起电话之后，推销员才会来跟你讲话。你一旦知道了这个状况，接了电话后若有暂停时间，等于是对方泄密，就给了你机会在推销员接电话之前先断线。而抽样调查的电话在你接起来的时候，就应该有个访问员在电话线另一端等候。

置信叙述说的是一个事实，显现出所有可能样本会发生的状况，我们用它来表达对一个样本的结果有多少信心。“95%的信心”代表“根据我们用的抽样方式，有95%的时候可以得到与真正值这么接近的结果。”以下是对于如何解读置信叙述的一些提示：

- 置信叙述的结论永远是针对总体而不是针对样本。我们确实知道样本中1523位成人的情况，因为盖洛普调查访问了他们。置信叙述是根据样本的结果来对“所有成人”这个总体做某种结论。
- 我们对总体所做的结论永远不会是完全确定的。盖洛普的样本有可能就是误差超过3个百分点的5%样本之一。
- 抽样调查可以选择95%以外的置信水平。较高的置信水平的代价，是较大的误差界限。对于同一个样本来说，99%的置信叙述

例6 你赞成赌博吗？

盖洛普对于赌博的调查，是以这样的问题开始的：

“首先要请问，一般来说，你赞成还是反对合法的赌博？”

除了我们已经探询过买彩券习惯的1523位成人(18岁以上)之外，盖洛普还抽了一个包含501位青少年(13—17岁)的随机样本。样本结果是：

成人：1523位当中，有959位赞成 $\hat{p} = 959/1523 = 0.63$

青少年：501位当中，有261位赞成 $\hat{p} = 261/501 = 0.52$

在报告了以上及其他一些结果之后，盖洛普还说：

“对同样大小的样本结果来说，我们有95%的信心做以下的声明，由于抽样或其他随机因素所造成的误差，对成人(18岁以上)来说应是正负3个百分点，对青少年(13—17岁)来说应是正负5个百分点。”

你一定想到了：青少年样本比较小，所以关于青少年的结论，误差界限就比较大。我们有95%信心，有47% (即52%减5%)到57% (即52%加5%)之间的青少年赞成合法赌博。



的误差界限, 就比 95% 置信叙述的要大。如果你只要有 95% 的信心就很满足了, 得到的回馈就是较小的误差界限。要记得我们的速算法算的是 95% 信心的误差界限。

- 报告误差界限时, 用 95% 的置信水平是很普遍的。如果一则新闻报道中只说明误差界限而没有置信水平, 把置信水平当作 95% 是很安全的做法。
- 想在同样的置信水平下, 要求较小的误差界限吗? 取个大一点的样本就成了。你应该还记得较大的样本有较小的变异性吧。只要你愿意付出取够大样本的代价, 就可以要求所需的小误差界限, 并且仍然维持高的置信水平。

从大总体抽样

盖洛普的 1 523 名成人的样本, 相当于在美国成人当中, 每 130 000 人抽出 1 人。而这 1 523 人是在总体当中每 100 人抽 1 人, 还是每 130 000 人抽 1 人, 有关系吗?

• 总体大小无所谓

从一个随机样本所得到的统计量的变异性, 并不受总体大小影响, 只要总体至少比样本大 100 倍即可。

对于从随机样本算出的统计量的表现, 为什么总体的大小影响很小呢? 请想像以下的状况: 假使我们要从已收获的玉米中抽样, 因此把勺子塞进玉米粒当中。勺子并不知道它是在一袋玉米当中, 还是在卡车的玉米当中。只要玉米混得很均匀(如此则勺子舀出的是随机样本), 所得结果的变异性就只与勺子的大小有关。

这对于像盖洛普这样的全国抽样调查是好消息。一个大小为 1 000 或 1 500 的随机样本, 因为样本够大, 所以有低变异性。但是要记得, 因为自发性回应样本或方便样本有偏差, 所以再大也没用。因此, 把样本加大并不能消除偏差。

然而, 样本统计量的变异性是由样本大小决定, 而不是由总体大小决定, 对任何计划在一所大学里或一个小城中做抽样调查的人来



统计学上的争议

选举民意调查该禁止吗？

选前民意调查(preelection poll)告诉我们，俄亥俄州的选民有 58% 支持某位参议员。媒体很欢迎这些民意调查，但统计学家可就不大喜欢了，因为即使调查过程完全使用正确的统计方法，实际投票结果也常和调查结果相左。接受过访问的人，有许多在选举前改变主意，有些说还没决定，还有些人会告诉你他支持谁，可是到选举时却根本不去投票。选举预测是抽样调查当中，结果较不理想的一种，因为我们必须“现在”问选民，“未来”他要投谁。

在投票者离开投票所时进行访问的出口调查(exit poll)，就没有上述的问题。样本里面的人，是刚刚才投过票的。好的出口调查是根据从全美国选区抽出的样本来做的，常可以在离投票结束还很久时，就准确预测出美国总统大选的结果。而选举预测的政治效应之争又因此更加激烈。

反对选前民意调查的言论指控民意调查会影响选民行为。如果民意调查预测的结果一面倒，选民可能决定留在家里不去投票了，因为既然已经有既定的结果，干嘛还那么麻烦去投票呢？出口调查尤其令人忧心，

因为这等于是在选举完成前就报道实际选举结果。美国电视网同意，在任一州的选举结束前，不在该州公开发布出口调查的结果。如果某次总统选举的得票有点差距的话，电视网可能在下午两三点时就知道结果了，但电视网只会将在各州的选举都一一结束时，才一州一州地公布调查结果。即使如此，总统选举的结果仍可能在西岸选举结束前(因时差的关系，东岸选举早已结束)就知道了。有些国家法律明文规定禁止选前预测。在法国，总统选举前一周内不能公布任何民意调查结果，加拿大禁止在全国选举前 72 小时内公布民意调查结果。总计约有 30 个国家，对公布选举民意调查的结果有限制。

赞成选前民意调查的论点很简单：不应禁止信息的公开。选民应该自己决定怎么用这些信息。毕竟，候选人如果落后很多，他的支持者不必靠民意调查也会知道。限制发表调查结果反倒鼓励非法。在法国，仍有候选人在选举前一周内私下做调查(当然比公开调查不可靠)，然后把结果透露给记者，试图影响媒体报道。



说,这可就是坏消息了。举例来说,不管是要估计俄亥俄州州立大学的学生中,在政治方面属于保守派的比例,还是要估计美国所有成年居民中的保守派比例,只要是要求同样的误差界限,就得要抽取一样大的 SRS。即使俄亥俄州州立大学只有 4.8 万名学生,而美国的成人超过 2 亿,也不代表在俄亥俄州州立大学可以取一个较小的 SRS。

网络寻奇

进入 <http://www.gallup.com/help/FAQs/poll.asp> 这个网址,你就可以读到盖洛普的官方解释,说明为什么对于一个很大的总体,只从一个小得出人意料之外的样本中,就可以得到可靠的结论。



本章重点摘要

抽样的目的，是要从样本撷取关于总体的信息。我们通常用样本**统计量**，来估计总体**参数**的值。本章说明了一个重大观念：要描述一个样本是否值得信任，只要问：“如果我们从同一个总体抽取很多个样本，会发生什么情况？”假设几乎所有样本得出的结果都接近真正的值，那么即使并不确定我们的样本是否接近真正的值，还是可以对这个样本有信心。

在策划一项抽样调查的时候，第一目标要减少**偏差**，方法是用随机样本，而避免像自发性回应这种坏抽样方法。其次，抽的样本要够大，才能减低结果的**变异性**。用够大的随机样本，就能保证几乎所有样本都能得出精确结果。要表达我们对总体所做结论的精确程度，可以用**置信叙述**。新闻报道中常常只提**误差界限**。所提出的误差界限，大部分时候是针对 95% **置信水平**而来的。也就是说，如果我们抽许多样本，则有 95% 的机会，总体的真正值会落在误差界限之内。对于大小为 n 的简单随机样本，在 95% 置信水平之下，我们可以用 $1/\sqrt{n}$ 这个公式来估计误差界限。这个公式似乎显示出，重要的是样本的大小，而不是总体的大小。且只要总体比样本大很多，这一项就永远为真。



第3章 习题

3.1—3.4 当中的每一个黑体数字，都是参数或者统计量的值。请指出每个黑体数字是**参数**还是**统计量**。

3.1 美国劳工部宣布，他们在上个月访问了 50 000 个住户的样本当中，所有属于劳动人口(labor force)的人；而被访问的人中有 **4.5%** 失业。

3.2 一整货车的球轴承，平均直径是 **2.503** 厘米，这在买主对整批货要求的接受范围之内。检查者从这批货中抽验 100 个球轴承，得到平均直径 **2.515** 厘米，这超过了要求的标准，所以整批球轴承就误被退货了。

3.3 洛杉矶一位电话推销员，利用一种随机数字拨号装置，来对该市的住宅电话随机拨号。在最先拨的 100 个号码当中，有 **43** 个是电话号码簿当中没有登记的。这倒不令人惊讶，因为洛杉矶的住宅电话有 **52%** 都没列在电话簿中。

3.4 选民登记的记录显示，印第安纳波利斯的选民中，有 **68%** 登记为共和党员。为了试用一个随机数字拨号装置，你用该装置拨了 150 个随机选出的印城住宅电话号码。在所联络到的登记选民当中，有 **73%** 是登记在案的共和党员。

3.5 抽样实验。图 3.1 和图 3.2 显示出，如果我们从同一总体抽取许多样本时，样本比例 \hat{p} 值的情况。你可以照着同样的步骤，执行一个小规模的实验。

图 3.4 当中是一个小型总体，其中每一个圆圈代表一个成人。有色的圆圈代表不赞成合法赌博的人，而白色的圆圈代表赞成的人。你可以检查一下，共 100 个圆圈当中，有 60 个是白色的，所以在这个总体当中，赞成赌博的比例是 $p = 60/100 = 0.6$ 。

(a) 圆圈上面有从 00、01 到 99 的代码。用表 A 的列 101 来抽大小为 5 的 SRS。你的样本当中，赞成赌博的人的比例 \hat{p} 是多少？

(b) 再取 9 个大小为 5 的 SRS(则总共有 10 个 SRS)，这次用表 A 的

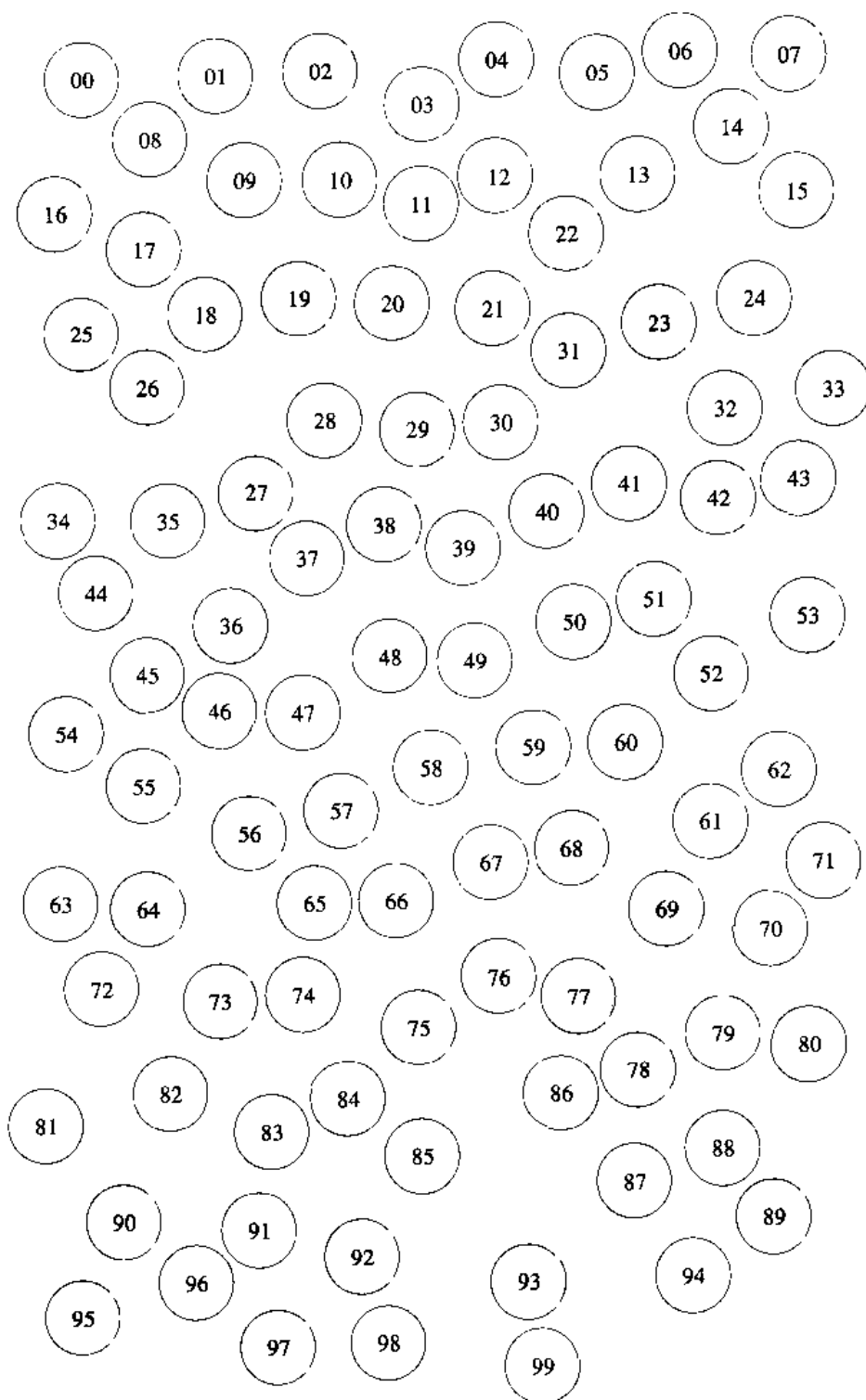


图 3.4 习题 3.5 的总体, 包含 100 个个体。有些个体(白色圆圈)赞成合法赌博, 其他的不赞成



列 102—列 110, 每个样本从新的一列开始。那你现在就有 10 个样本比例 \hat{p} 的值了。

- (c) 因为你的样本里面只有 5 个人, 所以 \hat{p} 的可能值只有 0/5、1/5、2/5、3/5、4/5 及 5/5, 也就是 \hat{p} 必定是 0、0.2、0.4、0.6、0.8 或 1 的其中之一。在一条直线上把这些数字标示出来, 并且替这 10 个结果造一个直方图, 做法是在每个数字上画一条垂直线段, 线段的长度就是结果等于该数字的样本个数。
- (d) 从一个大小为 100 的总体中抽取一个大小为 5 的样本, 当然不大符合一般实际的情况, 但不管怎样, 我们还是来看看你的结果。你的 10 个样本当中, 有几个把总体 $p = 0.6$ 估计得完全正确? 总体真正值 0.6, 对于你所有样本值来说, 是不是大致在中间的位置? 说明一下在取很多样本的时候, 为什么 0.6 会在所有样本值的中间位置。

3.6 抽样实验。我们用小总体当中的小样本, 来说明样本变异性。下列 25 位俱乐部会员当中, 有 10 位是女性。她们的名字旁边加了星号。俱乐部要随机选出 5 位会员, 提供他们去参加全国大会的免费旅程。

亚方索	达温	赫恩斯汀	迈杜	佛克特 *
比奈特 *	艾普斯汀	吉梅奈兹 *	斐瑞兹 *	温特
布鲁门巴	费瑞	鲁奥	斯班塞 *	威尔森
契丝 *	冈萨雷斯	牟欧 *	汤姆森	易克斯
陈 *	辜普塔 *	莫拉勒丝 *	涂明	钦莫

- (a) 抽取 20 个大小为 5 的 SRS, 每次都从表 A 的不同部分。把你每个样本当中的女性人数记录下来, 并画一个像图 3.1 里面的直方图来表达你的结果。在你的 20 个样本中, 女性人数平均是几人?
- (b) 如果 5 张免费票没有一张是给女性的, 你觉得俱乐部会员应不应该怀疑这其中存在性别歧视?

3.7 加拿大的全民医疗系统。加拿大安大略省的卫生部想要知道, 全民医疗系统在该省有没有达成应有的目标。有关医疗系统的信息, 大部分来自于病人的病历, 但是掌握该信息的单位不准我们用那些资



料, 来比较使用医疗系统和不使用医疗系统的人。所以卫生部就进行了一项安大略健康调查(Ontario Health Survey), 在住在安大略省的人当中, 访问了 61 239 人的随机样本。

(a) 这项抽样调查的总体是什么? 样本是什么?

(b) 调查发现在过去一年当中, 样本中有 76% 的男性及 86% 的女性, 至少去看了一次家庭医生。你认为这些估计值会不会接近整个总体的真正值? 为什么?

3.8 取大一点的样本 在抽样调查中用大些的随机样本有什么好处, 请用你自己的话来说明。

3.9 抽样变异性 在讨论盖洛普大小为 1 523 的样本时, 我们曾经问过这个问题: “可不可能有一个随机样本说 57% 的美国成人最近买过彩券, 而另一个随机样本却说只有 37% 的人买过呢?” 现在看一下图 3.2, 这里的直方图显示的是, 当总体真正比例是 $p=0.6$, 也就是 60% 的时候, 由 1 000 个大小为 1 523 的样本中得到的结果所画出的分布情况。如果从这个总体抽出的一个样本的结果是 57%, 你会不会觉得惊讶? 如果有个样本的结果是 37%, 你会不会惊讶呢?

3.10 平衡预算 一项访问了 1 190 位美国成人的调查, 显示有 702 位宁愿政府能够平衡预算而不是减税。这项结果的误差界限, 经报道为正负 4 个百分点, 但是报道中没说置信水平。不过, 我们可以相当有把握, 它应该是 95%。

(a) 宁愿平衡预算的人的样本比例 \hat{p} 的值是多少? 请叙述说明, 这题里面的总体比例 p 指的是什么?

(b) 对参数 p 做一个置信叙述。

3.11 偏差及变异性 图 3.5 中的直方图, 表达出在 4 个不同情况下, 抽许多样本所得到样本统计量的值, 其分布情况如何。这些图类似于图 3.1 和图 3.2 中的直方图。也就是说, 长方形的高度代表的是, 从同一总体抽许多样本时, 有多少个样本的样本统计量值, 会落在那个长方形的底部范围内。总体参数的真正值也标明在图上。把图 3.5 中的每个图, 依高偏差或低偏差, 以及高变异性或低变异性归类。

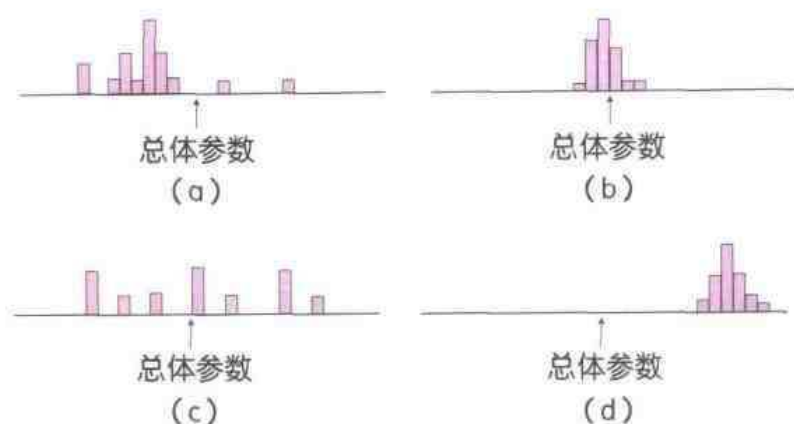


图 3.5 从同一总体抽取许多样本, 并根据一个样本统计量对不同样本所得到的值, 所得到的直方图。这 4 个图是针对 4 个不同抽样方法所得的结果, 相关问题在习题 3.11

3.12 预测选举结果. 就在一次美国总统选举之前, 一项全国性的民意调查, 把他们每周抽样的样本大小, 从通常的 1 500 人增加到 4 000 人。这个比较大的随机样本是否会把调查结果的偏差减低? 是否会把调查结果的变异性减低?

3.13 取大一点的样本。一个学管理的学生正在计划做一份报告, 主题是大学生对于开学期间打工的看法。她设计好一份问卷, 并计划要随机选取 25 位学生来填答问卷。她的导师认可了她的问卷, 但建议她把样本大小增加到至少 100。为什么大一点的样本比较好? 用速算法分别计算样本大小为 25 及样本大小为 100 时的误差界限, 来支持你的说法。

3.14 在美国各州抽样 一个美国联邦政府的委托单位计划要在每一州居民当中抽取 SRS, 来估计每一个州里面拥有房地产的居民比例。每州居民人口最少的是怀俄明州的 525 000 人, 最多的是加州的 3 300 万人。

- (a) 如果在每一州取一个大小为 2 000 的 SRS, 样本比例的变异性, 会不会各州不同? 要说明你的答案。
- (b) 如果在每一州取全州人口的 $1/10$ 个百分点 (0.001) 的 SRS, 则样本比例的变异性会不会各州不同? 要说明你的答案。



3.15 对女性做意见调查。《纽约时报》为了对某些女性议题做民意调查，从美国全国(阿拉斯加和夏威夷除外)随机抽取了1025位女性来访问。调查发现，有47%的女性说，她们没有足够的个人时间。

(a) 调查结果说，在95%的置信水平之下，误差界限为 ± 3 个百分点。对于在所有女性当中，觉得个人时间不够的人的比例，做一个95%的置信叙述

(b) 向某个完全不懂统计的人解释，为什么我们不能只是说“全部女性当中有47%觉得个人时间不够。”

(c) 然后解释明白，95%置信水平是什么意思。

3.16 哈里斯(Harris)调查。哈里斯调查在解释他们结果的精确程度时，用了以下的说法：“理论上来说，对于这样大小的样本，我们可以95%确定地说，调查所得结果、和假如用全部的总体所得的完全精确的结果比起来，其统计精确性(statistical precision)是正负3个百分点。”哈里斯调查说的“95%确定”是什么意思？

3.17 对男性和女性做抽样调查。习题3.15中描述的抽样调查，除了1025位女性之外，也访问了472位随机选出的男性。调查报告当中对于女性的结论，宣称在95%置信水平下，误差界限是 $\pm 3\%$ 。而对于男性的结论，误差界限是 $\pm 5\%$ 。为什么这比女性的误差界限大？

3.18 解释置信水平。一位学生读到以下叙述，我们有95%的信心，美国年轻人在“全国教育进展评估”(National Assessment of Educational Progress)中数量部分(quantitative part)的平均分数，会在267.8—276.2之间。有人要求这位学生说明这段叙述的意义，学生回答：“所有年轻人当中，有95%的人所得分数在267.8—276.2之间。”他说得对不对？请说明你的答案。

3.19 枪支暴力。哈里斯调查访问了1009个成人的样本，问他们认为将来哪种死因会更普遍，答案当中名列第一的是枪支暴力，有70%的人认为，因为受枪击而死亡的人数会增加。

(a) 受访的1009人当中，有多少人认为因枪支暴力而死亡的人数会增加？

(b) 哈里斯说这次调查的误差界限是正负3个百分点。说明如何向一个不懂统计的人解释“误差界限是正负3个百分点”的意义。



3.20 算出误差界限 例6告诉我们,盖洛普问了501位青少年是否赞成合法赌博;52%说赞成。用速算法估计一下,对所有青少年下结论时的误差界限是多少?你的结果和例6当中盖洛普的误差界限比较起来有何差别?

3.21 算出误差界限 习题3.19考虑的是哈里斯调查1009人的样本。用速算法估计,如果对所有成人做结论,误差界限会是多少?你的结果和哈里斯宣布的3%误差界限接近吗?

3.22 算出误差界限 习题3.7里谈到了一项对于住在安大略省61239位成人的抽样调查。若要对安大略省全体成人做结论,在95%置信水平之下,误差界限大约是多少?

3.23 有没有地狱存在?某篇新闻报道说,在最近做的一项民意调查中显示,样本里的1108位成人当中,有78%说他们相信有天堂,而只有60%的人相信有地狱。

(a) 用速算法估计一下,这样大小的样本,误差界限是多少。

(b) 对所有成人当中相信有地狱的比例,做一个置信叙述。

3.24 害怕遇到歹徒 盖洛普问一个1493位成人的随机样本:

“你是否因为害怕遇上歹徒,所以在夜里就算只出门不到1英里也不敢?”样本当中有672个人回答:“是的。”对所有成人当中,因为怕遇上歹徒而不敢夜里出门的比例,做一个置信叙述。(用速算法求误差界限。)

3.25 较小的误差界限 习题3.21中的民意调查,访问了1009位成人。假设你希望误差界限,只有那题所算出的一半,那你计划要访问多少人才够?

3.26 取悦国会 习题3.10里谈到一项对1190位成人做的抽样调查,对应95%的置信水平,误差界限是 $\pm 4\%$ 。

(a) 有位美国国会议员认为95%的置信水平不够。他希望可以99%的信心。对同一个样本来说,99%信心的误差界限和95%信心的误差界限,有何差别?

(b) 另一位国会议员觉得95%的信心够好了,但是她想要比 $\pm 4\%$ 小



的误差界限。我们怎样可以维持 95% 置信水平，且得到较小的误差界限？

3.27 当前人口调查：虽然民意调查通常都使用 95% 置信叙述，但还是有抽样调查是用其他的置信水平的。举例来说，美国每月失业率是根据当前人口调查的约 50 000 住户得来的。随着失业率一起公布的误差界限，是大约 0.2 个百分点，置信水平 90%。相较之下，95% 置信水平的误差界限会比较小还是比较大？为什么？

3.28 千禧年的乐观展望：在 2000 年 1 月的时候，一项盖洛普调查询问了一个 1 633 位美国成人的随机样本，问道：“大体来说，你对目前美国的状况是否满意？”结果有 1 127 位说满意。针对这项发现写一个简短的报告，可以假想你是在替报纸写报道，且不要忘了误差界限。

3.29 模拟(Simulation)：随机数字可以用来模拟随机抽样的结果。假设你要从一个包含许多大学生的大总体里面，抽取一个大小为 25 的简单随机样本，总体当中有 20% 的学生，暑假没在工作。要用随机数字模拟这个 SRS 的话，我们可以令表 A 当中任一处开始的连续 25 个数字，代表我们抽取的样本当中的学生。让 0 和 1 这两个数字代表没工作的学生，其他数字代表有工作的学生。这样的设计，是对我们要的 SRS 的一个正确的模拟，因为 0 和 1 两个数字，在所有 10 个被选中概率相同的数字当中占 20%。

照以下步骤来模拟 50 个随机样本的结果，把表 A 共 50 列里面，每一列的头 25 个数字当做一个样本，数一数每一个样本当中 0 和 1 总共有几个。把 50 个样本的结果，用像图 3.1 一样的直方图展示出来。总体的真正值(也就是未工作的比例，20%)，是不是靠近你的图的中央？在你的 50 个样本当中，未工作的学生人数，最大是多少，最小是多少？在你的样本里面有 4、5 或 6 个学生未在工作的，占你 50 个样本的多少百分比？

第 4 章

真实世界中的抽样调查

民意调查面面观

某项民意调查访问了随机选出的 1 000 位群众后，公布调查结果，其中包括误差界限。我们是不是就应该满意了呢？恐怕不行。有许多调查并没有把和样本相关的信息全盘告诉我们。普优研究中心 (Pew Research Center) 模仿了几家较好的民意调查机构的做法，然后把过程细节详述如下。

大部分民意调查是利用电话进行的，方法是用随机拨号的方式来取得住宅的随机样本。在剔除掉传真机号码以及公私营机构的号码之后，普优总共必须打 2 879 个住宅电话，才能得到 1 000 人的样本。这 2 879 个电话可以分类成以下几种情况：



从来没人接电话	938
接了电话却拒绝接受访问	678
条件不合：没有 18 岁以上的人，或语言不通	221
访问未完成	42
访问完成	1 000
总计电话数	2 879

在 2 879 个有效的住宅电话当中，有 33% 从来没有人接。其他接了的人里面，有 35% 不愿接受访问。整体的无回应(*nonresponse*)比例(这些包括不接电话、不愿接受访问、没有完成访问的人)，占 2 879 人当中的 1 658 人，也就是 58%。普优在 5 天的期间内，选一星期中不同的日子，和每天不同的时段，每个号码都打了 5 次。很多调查都只打一次电话，而且常常在接到电话的人当中，有超过一半不愿接受访问。在现实世界当中，简单随机样本一点也不简单，而且也不一定随机。这就是本章的主题。

抽样调查怎样出错

随机抽样方法在选样本时可以消除偏差，也有办法可以控制变异性的_{大小}。所以是不是只要我们有看到“随机选取”和“误差界限”这两个关键词时，就可以信任眼前的信息了呢？它当然是好于自发性回应，但是有没有像我们希望的那么好，可就不一定。在真实世界里抽样，比起从教科书习题里的名单当中抽一个 SRS，要复杂得多，结果也较不可靠。置信叙述并没有把真实抽样的所有误差来源都反映出来。



• 抽样会发生的误差

抽样误差(sampling error)是抽样这个动作所造成的误差。抽样误差使得样本结果和普查结果不同。

随机抽样误差(random sampling error)是样本统计量和总体参数之间的差距,是在选取样本时因机遇造成的。置信叙述中的误差界限只包含随机抽样误差。

非抽样误差(nonsampling error)是和“从总体取样本”这个动作无关的误差。非抽样误差即使在人口普查中也可能出现。

大部分的抽样调查,都会遇到随机抽样误差以外的误差。这些误差可能导致产生偏差,使得置信叙述变得没有意义。好的抽样技巧都有减少各种误差来源的技术。这种技术有一部分是统计科学,因为随机样本及置信叙述都属于统计科学的范畴。然而实际应用上,要得到好的样本,光靠好的统计是不够的。我们来看看抽样调查有些什么样的误差来源,以及抽样者如何与之奋斗。

抽样误差

随机抽样误差是抽样误差的一种。误差界限告诉我们随机抽样误差的严重程度,而我们可通过选择随机样本的大小,来控制随机抽样误差。另一个抽样误差来源是使用了糟糕的抽样方法,比如自发性回应。糟糕的方法是可以避免的,但其他的抽样误差可就没那么好对付了。抽样之前必先要有一个“清单”,上面列出总体所有成员,可让我们从中抽取样本。我们称之为**抽样框**(sampling frame)。理论上来说,抽样框应该包括总体当中的每一个个体。但是整个总体的清单通常很难取得,所以大部分的样本,多多少少都会有涵盖不全的问题。

如果抽样框原本就漏掉了某些群体,那么即使从这个抽样框当中抽取随机样本,所得结果还是有偏的。比如说,假如我们用电话号

• 涵盖不全

在选样本的过程中,如果总体当中的有些部分,根本未被纳入选择范围,这时就发生了**涵盖不全**(undercoverage)的问题。



码簿当做电话访问的抽样框，我们会漏掉所有未将电话号码登记于电话簿的人。在很多大城市里，有超过一半的住户电话没有登记，所以抽样调查的结果，会对城市居民有较大的涵盖不全问题和较大偏差。不过，事实上电话调查是利用随机数字拨号系统，在选定的访问区域中随机拨出电话号码。这样做的效果，等于是把所有住宅电话都纳入了抽样框。



虽然退出社会活动有时会觉得无聊，不过只要能增加意见调查的涵盖不全，泰德就觉得一切都值得了

例1 我们的确涵盖不全

大部分的民意调查无力去试图涵盖全美国成年居民这样大的总体。且他们是用电话做访问的，因此会漏掉没装电话的那6%住户。而他们只联络一般住户，所以住在宿舍的学生、监狱里的犯人以及大部分的军人都被排除在外；而且还漏掉了无家可归的人和住在临时收容所的人。另外因为打电话到阿拉斯加和夏威夷很贵，所以大部分民意调查的取样，并不包括这两州在内。还有，很多民意调查只用英语访问，这又把某些移民家庭给排除在外了。



出现在大部分抽样调查中的涵盖不全类型，最容易漏掉年轻人、穷人或常常搬家的人。不过随机拨号系统所产生的样本，可以说很接近有电话住户的随机样本，但这并不包含阿拉斯加和夏威夷地区。在谨慎执行的抽样调查中，抽样误差通常不大。真正的问题是在有人接电话(或不接电话)时开始。现在非抽样误差登场了。

非抽样误差

非抽样误差是连人口普查都可能逃不过的差错。非抽样误差包括处理误差(processing error)，也就是在机械化工作时犯的错误，例如计算错误或将受访者的回答输入电脑时犯的错误。而电脑的普及使得处理误差比以前少。

另一种非抽样误差是回应误差(response error)，这在受访对象给了不正确回答时就会发生。受访对象也许会谎报年龄或收入，或者对于是否曾用过禁药没有诚实答复。在被问到上星期共抽几包烟时，她也可能记错答案。受访对象也许没听懂问题，但宁愿猜，也不愿显得

例 2 电脑辅助访问

访问员手上拿着写字板的日子已经成为过去，现在的访问员都是以电脑辅助访问(computer-assisted interviewing)，他们若不是带着笔记本电脑去做面对面访谈，就是一边看着电脑屏幕一边做电话访问。访谈由电脑软件来处理，访问员从电脑屏幕上读取问题，再用键盘把回答输入。电脑自动跳过不相干的问题，例如只要受访者说没有小孩，后续关于小孩的问题就不会出现。电脑可以检查有关问题的答案是否一致，还可以随机排列问题顺序，以免老用相同顺序问问题而造成偏差。

电脑软件也可以处理抽样的过程。电脑软件会记录哪些人已经回答了，而且把这些回答存档。以前要把答案从纸上转入电脑是很繁琐的工作，也是处理误差的一大来源，而现在这些已成为历史。电脑甚至可以安排电话调查的打电话时间，并会考虑受访对象所在的时区，如果有人第一次接电话时表示有意愿但没时间回答而重约时间，电脑也会履行这个约定。



无知。若问到关于受访对象在一段固定时间内的行为，尤其容易因记忆错误而导致回应误差。比如说，美国“全国健康调查”(National Health Survey)问大家去年总共看了几次病，结果对照健康记录之后却发现，人们会把看病的次数忘掉60%。而有关敏感议题的问题也容易发生回应误差，从以下的例子就可看出。

例3 种族效应

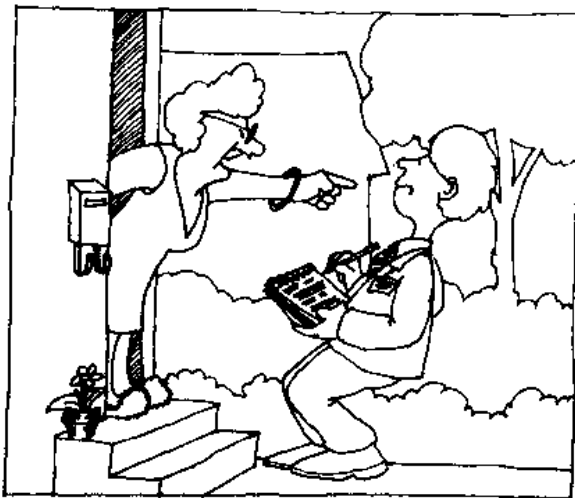
1989年，纽约市选出第一位黑人市长，维吉尼亚州选出第一位黑人州长。这两个事件，在投票所访问投完票的选民时，所预测到的胜负差距，都比实际开票的差距大。因此调查机构相当确定，有些受访选民因为不愿承认没投票给黑人候选人而说了谎。

• 无回应

无回应(nonresponse)是无法得到已经被选入样本中的个体的资料。最常发生无回应的原因，是联络不上受访对象或受访对象拒绝合作。

利用现代科技再加上注重细节，可以把处理误差减到最低。技巧熟练的访问员也可以大幅度的减少回应误差，特别是在面对面访问的时候。但是对于像无回应这种最严重的非抽样误差，并没有简单的方法可以对付。

无回应是抽样调查所面临最严重的问题。人们愈来愈不愿意回答问题，尤其是在电话上。电话推销、答录机以及来电显示日渐普遍，也减低了对于电话调查的回应比例。封闭的社区和有警卫的大楼，又阻碍了面对面的访问。无回应会使调查结果有偏差，因为不同群体的人有不同的不回应率。比如说，老人和大城市居民的拒答率就比较高。无回应造成的偏



“你可以打电话，送电子邮件，站在门口一整天，但答案还是‘不’！”



差，很容易就可超越误差界限所描述的随机抽样误差。

做抽样调查的人知道一些减低不回应率的技巧。只要对方肯接电话，受过严格训练的访问员就有办法让他们不挂掉。而没访问成功的，隔久一点再打回去也有用，或者在打电话之前，先寄封信也有帮助。但是又要寄信，又要不断打电话回去，会使得调查进度缓慢，所以需要很快得到答案来满足媒体的民意调查，就不会使用这些方法。即便是过程最严谨的调查，也仍然被不回应所困扰，而且不论如何专业，也无法完全克服这个难题。所以下面的这段叮咛，就更显得重要。

例 4 无回应糟糕到什么程度？

“当前人口调查”(CPS)是我所知道的美国的所有调查当中，回应率(response rate)最高的：CPS 样本当中，只有 6% 或 7% 的住宅不予回应。人们比较会对像 CPS 这样的政府调查做回应，而且 CPS 会先登门拜访他们的样本，之后才用电话做访问。

“全面社会调查”(见第 1 章例 6)也会和样本面对面，而且这项调查是由一所大学来做的。即使占有这些优势，最近做的调查仍有 24% 的不回应率(rate of non-response)。

而媒体、市场调查或者民意调查公司所做的民意调查又如何呢？我们不知道他们的不回应率，因为他们不肯说，但是这就表示不妙。从我们之前见过普优研究中心所提出的数字，可以看出情况有多么糟糕。普优共得到 1 221 个回应(其中有 1 000 人属于他们锁定的总体)，以及总共 1 658 次调查主人不在家、拒绝受访或者没有访问完。这样算起来的不回应率，是 2 879 当中占 1 658，也就是 58%。普优的研究员已经比许多民调的调查员做得彻底。有内部人士透露，民意调查的不回应率，常常达到初始样本的 75% 或 80%。

• 误差界限不包括什么？

一项抽样调查中所宣布的误差界限，只包括随机抽样误差。涵盖不全、无回应以及其他实际困难也会造成大偏差，但是误差界限并没有包含这些项目在内。



严谨的抽样调查会告诉我们这些真相。比如盖洛普调查就会宣称：“除了抽样误差以外，问题的措辞以及执行调查时遇到的实际困难，会导致民意调查结果有偏差或产生其他误差。”说得真对！

无回应是不是使得很多调查结果都没用了呢？也未必如此。本章开始的时候，我们说明了普优研究中心执行的一项“标准”电话调查。普优研究中心其实也执行了一项“严格”调查：在打电话之前先寄信，然后在8周内不断打电话，再寄快递信件给拒绝受访的人，等等。所有这些做法，把不回应率降到了30%，而标准调查的不回应率是58%。

然后普优比较了两项调查中，对同样问题的答案。两个样本在年龄、性别及种族各方面都相当接近，只不过严格样本比较富一点。两个样本也都对除种族以外的所有议题，有接近的看法。而起先不肯接受访问的人，对黑人的同情比较低些。整体来说，标准调查所得的结果，其精确程度在合理的范围。不过就像例3里的状况一样，种族议题仍然是例外。

问题的措辞

最后一项对抽样调查结果有影响的是问题的措辞。要把问题表达得完全清楚是出乎意料的困难。有个调查问到：“是否拥有‘stock’（股票，也是家畜）？”大部分的德州牧场主人都答：“是”，可是他们拥

例5 稍改几个字，结果就大不同

美国人对于政府对穷人的帮助，看法如何？只有13%的成人认为他们花太多钱在“帮助穷人”上，可是却有44%的成人认为他们花太多钱在“社会福利”上。苏格兰人对于从英国独立出来，看法如何？有51%的人赞成“苏格兰独立”，但是只有34%的人支持“从联合王国(United Kingdom)分离出来的独立的苏格兰”。

好像“帮助穷人”和“独立”是好的字眼，令人充满希望，而“社会福利”和“分离”就是负面的字眼。只把问题的措辞稍改一点，回答就有可能大幅改变。



有的，大概不是在纽约证券交易所可以买卖的那种。

问题的措辞总是会影响答案。如果问题的说法倾向于某个答案，则又是一个非抽样误差的来源。有一招受欢迎的把戏是问受访者是否赞同某项政策以便达到某种目标：“你是否赞成禁止私人拥有枪械以减低犯罪率？”及“你是否赞成可以判死刑以减低暴力犯罪的比例？”都是“加了料”的问题，可以让担心犯罪率的人都给予肯定的答复。下面的例子就是有引导倾向的问题造成的影响。

例 6 选举经费补助

政治活动的经费补助总是争议不断。以下是意见调查问卷上关于这个议题的两个问题。

是否应该立法消除所有可能的途径，使特殊利益团体无法捐献大笔款项给候选人？

应该立法来禁止利益团体捐助竞选活动？或者团体有权捐款给所支持的候选人吗？

第一个问题是佩罗(Ross Perot)提出的，佩罗是美国1992年总统选举时的第二党候选人。邮寄来的回答中，99%答“是”。我们知道自发性回应的调查结果是没用的，所以杨克洛维奇(Yankelovich)调查公司对全国随机样本问了同样的问题，结果80%答“是”。佩罗的问题简直是要求人家答“是”，所以杨克洛维奇写了第二个问题，用较中立的立场来提出这个议题，在问这个问题时，全国随机样本中只有40%赞成禁止捐款。

如何应对非抽样误差

非抽样误差，尤其是无回应，是躲也躲不掉的。严谨的抽样调查



应该要怎样处理这些问题呢?首先,用其他住户来取代不回应的人,因为城市里的不回应率比较高,如果用不回应住户附近的其他住户来取代,可以减低偏差。一旦数据搜集完成之后,所有专业的调查机构都会用统计方法来给回应加权,以期导正偏差来源。如果城市里太多住户没回应,就给城市里有回应的那些结果加权。如果样本里太多女性,就给男士们加权。举例来说,以下就是《纽约时报》对它某次抽样调查的部分描述:

考虑到每个住户人数和电话部数都有不同,也为了对样本中有关地理位置、性别、种族、年龄以及教育程度的各种差异做调整,此调查结果已经过加权处理。

其目标是要使得结果“好像”是从一个在年龄、性别、住户地理位置以及其他各种变量都和总体符合的样本得知。

执行加权这件事,替统计学家制造了许多工作机会。而这也表示,抽样调查所宣布的结果,极少是像表面上看起来那么简单。盖洛普宣布,他们访问了1523位美国成人,发现其中有57%在过去12个月当中买过乐透彩券。从表面上看起来,因为1523的57%是868,所以盖洛普的样本当中,应该是有868个人玩彩券。然而事实却非如此。盖洛普无疑曾用了某些特殊的统计技巧,来给实际得到的结果加权;57%这个数字,是盖洛普对于如果没有人不回应时,所应该得到的结果的最佳估计。加权的确可以修正偏差,但也通常会增加变异性。在宣布误差界限之前必须把这些都考虑进去,这又给了统计学家更多的工作机会。

真实世界中的抽样设计

抽样的基本概念很清楚:从总体抽一个SRS,用一个从你这个样本得来的统计量,估计某个总体参数。现在我们已经知道,为了能够对无回应问题做补救,样本统计量被人在背后“动过手脚”。统计学家也对我们钟爱的SRS“伸出魔手”。在真实世界中,大部分抽样调查使用的是比SRS还复杂的抽样设计。



他先动手的!

有人研究酒吧里的打架致死事件,发现其中有90%都是死掉的那个人先动手的。这种结果你可别相信。假如你跟人打架把人给揍死了,警察问你谁先动手的时候,你会怎么回答?反正死人也不会说话。这也是无回应的一种。

例7 当前人口调查(CPS)

CPS的总体,包括美国的所有住户(阿拉斯加及夏威夷也在内)。样本是分阶段抽的。普查局把全美国分成2007个地理区,称之为基本抽样单位(PSU, Primary Sampling Unit)。大体来说,是把邻近的县组成一个单位。第一阶段抽样抽出792个PSU。不过这并不是SRS,因为如果所有的PSU被抽中的机会都是一样的话,样本里面很可能漏掉芝加哥和洛杉矶。所以有432个人口稠密的PSU,会优先收进样本。剩下1575个基本抽样单位,根据某些准则,把类似的结合为一个层(stratum),共形成360层。然后在每一层中随机选取一个PSU作为那层的代表。

第一阶段抽样所得的792个PSU再细分成普查街区(census block),是比较小的地理区域。普查街区再依住户种类、弱势群体等条件而分层。同一个普查街区的住户依地理位置排序,然后每约4户成一群(cluster)。最终取得的样本,是从街区的每一层抽出的群样本,而不是住户样本。访问员会去被抽出的群中的每一户访问。从每一个街区层(stratum of blocks)里抽出的群样本也不是SRS。为了确保选出的群在地理位置上能够适度地数开,样本是先随机选一个群,之后就抽取这个群后面的第10个(比如说)、第20个,等等。

CPS的统计,说明了在真实生活中,使用面对而访问的样本所共有的一些特质。先把住户组合成基本抽样单位,再集成成群,然后分阶段抽样,而且最后抽出的是群,这样的做法可以省掉访问员许多的旅行时间。例7中提到的各种细分概念中,最重要的是分层抽样(stratified sampling)。

要选择适当的“层”,当然必须根据抽样前对总体的了解才办得到。你或许会把大学里的学生,依大学部的或者研究所的分成两层,也可能依住校的和不住校的分成两层。分层样本有几点优于SRS:首



先，因为是在每层分别取 SRS，我们可以在每层决定样本大小，因此可以得到有关各层的个别信息。其次，分层样本的误差界限，通常比同样大小的 SRS 要小。理由是因为同一层中的个体之间，相似程度比起整个总体的个体之间来得大，所以借助分层考虑，可以消除样本中的某些变异性。

• 误差界限不包括什么？

选取分层随机样本有以下步骤：

步骤 1：将抽样框中的个体先分成若干群，叫做层。分层的标准是，你对于这些层有特别的兴趣，或者同一层中的个体有接近的性质。

步骤 2：每层各取一个 SRS，全部合起来就是我们要的样本。

例 8 把学生分层

一所大型的大学有 30 000 个学生，其中 3 000 个是研究生。如果抽一个 500 名学生的 SRS，每个学生被抽中的概率是相同的，概率是：

$$\frac{500}{30\,000} = \frac{1}{60}$$

我们预期在 SRS 中大约只有 50 个研究生，因为全部学生中 10% 是研究生，所以我们期望 SRS 中约 10% 是研究生。因为大小为 50 的样本不够大，无法适当精确地估计研究生的意见。不如用包含 200 名研究生及 300 名大学生的分层样本反而比较好。

你应该知道怎样选这样的分层样本吧？给研究生 0001—3000 的代码，然后用表 A 选择大小为 200 的 SRS。再给大学生 00001—27000 的代码，再用表 A 选出大小为 300 的 SRS。将这两个 SRS 摆在一起就是所要的分层样本了。

在这个分层样本中，每个研究生被抽中的概率是：

$$\frac{200}{3\,000} = \frac{1}{15}$$

大学生的机会就小多了，每人是：

$$\frac{300}{27\,000} = \frac{1}{90}$$



因为我们有二个 SRS, 就很容易可以分别估计大学生和研究生的意见。速算法告诉我们, 样本比例的误差界限, 对研究生来说大约应该是:

$$-\frac{1}{\sqrt{200}} = 0.07 \text{ (即 7\%)}$$

对大学生来说大约应该是:

$$\frac{1}{\sqrt{300}} = 0.058 \text{ (即 5.8\%)}$$

现在如果告诉你一件事, 可能会让你很惊讶: 分层样本可能违反了 SRS 最吸引人的性质之一, 也就是分层样本未必会给总体中每个个体同样被抽中的机会, 因为有些层在样本中所占比例有可能被刻意提高。

因为例 8 中的样本, 刻意加重了研究生代表, 所以最后的分析必

例 9 电话样本的苦恼

理论上来说, 利用随机拨号来进行的电话访问, 用 SRS 应该就可以了。用电话访问不太需要做群, 但分层可以减低变异性, 所以电话访问常分两阶段抽样: 先抽一个电话前码(区号加上电话号码前 3 位)的分层样本, 再在每个抽中的前码当中, 随机拨个别号码(最后 4 位)。

电话号码的 SRS 的真正问题是, 太少电话号码真正属于住户。这只有怪科技了。传真机、数据机以及移动电话用掉许多电话号码。在 1988—1999 年间, 美国的住户增加了 11%, 但是看起来像是住户电话的号码却增加了 90%。而且许多新的前码之下, 根本都还没有任何住户电话。电话访问现在利用“对照清单样本”, 先对照电子电话簿, 把底下没列电话号码的前码去掉, 剩下的前码才做随机抽样。这样可以减少白打的电话数, 但是如果有人位于所有电话都没登录在电话簿的地区, 就会被排除在外了。因此他们会对底下没列电话号码的前码另外再抽样(还是分层), 以弥补不足。



须做调整，才能得到所有学生意见的无偏估计。请记得我们的速算法只能用在 SRS 上。事实上，要做专业分析的话，还得特别考虑到总体“只有”30 000 个个体的这项事实，所以统计学家又有更多工作机会了。

从例 7、8 及 9 应该清楚地看出，设计抽样是专家的工作，连大部分统计学家也都无法胜任。所以我们不用花精力在这些抽样细节上，只要了解主要的概念是，好的抽样设计利用机遇从总体抽取个体。也就是说，所有好的样本都是概率样本(probability sample)。

• 概率样本

概率样本(probability sample)是利用机遇抽取的样本。我们要先知道哪些样本是可能的，以及每个可能的样本被抽中的概率是多少。有些概率样本，比如说分层样本，并不包括总体的所有可能样本，即使包括在内的样本，被抽中的概率也未必一样。

举例来说，一个含 300 名大学生及 200 名研究生的分层样本，必定是由 300 名大学生及 200 名研究生组成的；而 SRS 则可以由任何 500 名学生组成。这两种都属于概率样本。我们只需要知道，从概率样本得到的估计值，也拥有和从 SRS 样本得到的估计值同样好的性质。我们一样可以做出不偏的置信叙述，如果把样本加大，也一样可以缩小误差界限。像自发性回应样本这种非概率样本就没有这些优点，也无法提供值得信任的总体信息。现在我们已经知道，大部分全国性的样本比 SRS 更复杂，不过我们常常假装好的样本就是 SRS。这样既可以保留重要概念，也可以隐藏啰里啰唆的细节。

相信调查结果之前该问的问题

如果调查者使用好的统计技巧，而且认真准备抽样框，注意问题的措辞，并减少无回应，意见调查及其他抽样调查是可以提供精确且有用的信息。可是，很多调查，尤其是那些设计好要影响公众意见而不只是要记录意见的调查，并不能提供精确而有用的信息。在你留意一些调查结果以前，应该先问以下问题：

纽约，纽约

他们说纽约市更大、更富、更快、更粗鲁了。这可能不是随便说说而已。专业抽样调查公司国际佐格比(Zogby International)说，以全美国平均来讲，每打 5 个电话才能找到一个真的人说话。但是如果打到纽约，则要打 12 个才找得到。调查公司都派最好的访问员打电话到纽约，而且常常因为他们必须面对的压力而给付额外的红利。



- **谁做的调查?**就算是政党,也应该请专业的抽样调查机构来做,专业机构为了名声,会好好做调查。
- **总体是什么?**也就是说,调查是在寻求哪些人的意见?
- **样本是怎样选取的?**注意看他们有没有提随机抽样。
- **样本多大?**最好除了样本大小以外还有精确度的评估,像是所有用同样方法可能得到的样本中的 95% 会落进去的误差界限。
- **回应率是多少?**也就是说,原来预定的受访对象中有百分之多少确实提供了信息?
- **用什么方式联络受访者?**电话?邮寄?面对面访谈?
- **调查是什么时候做的?**是不是刚好在一个可能影响结果的事件发生之后?
- **问题确实是怎么问的?**

学术界的调查中心和美国政府的统计部门,都会在公布抽样调查结果时回答以上的问题。全国性的民意调查通常不宣布回应率(常常都很低),但会提供其他信息。编辑和新闻播报员则有坏习惯,常要删掉这些“无聊”内容而只报告结果。利益团体、地方报纸及电视台的许多抽样调查,则根本不理会这些问题,因为他们的调查方法事实上是不可靠的。假如有从政者、广告商或地方电视台宣布民意调查的结果,却没提供完整的信息,我们最好是抱持怀疑的态度。

网络寻奇

本章一开始的时候,将普优研究中心为人民与报刊组织做的一项研究介绍给读者,该中心的网址是 www.people-press.org。该网址中有各项报告,比如我们谈到这个研究的结果报告就在其中,以及该中心自己做的民意调查的结果。也许你有兴趣看看“民意调查分析”(Poll Analysis)那一节,里面对于目前的进行民意调查,包括普优做的,盖洛普等其他单位做的,对于社会大众对重要议题的看法提供了哪些信息,有很详细的评论。



本章重点摘要

即使是专业的抽样调查，也没法对总体提供完全正确的信息，因为抽样时会有许多可能的误差来源。抽样调查结果中提出的误差界限，只涵盖**随机抽样误差**，也就是在选随机样本时，因机遇而产生的变异。其他种类的误差没有被包括在内，而且也没法直接度量。**抽样误差**是由抽样这个行为造成的误差。随机抽样误差及**涵盖不全是抽样误差**中常见的两种。当总体中有些成员没被列进抽样框的时候，就发生涵盖不全的问题，**抽样框**是总体全部成员的清单，样本就是从这当中抽取的。

然而在大部分严谨执行的调查中，最严重的误差是**非抽样误差**。这些误差和抽取样本没有关系，连普查时也会有非抽样误差的存在。抽样调查最大的一个问题就是**无回应**：调查对象联络不上，或者拒绝回答。处理资料时发生的错误(**处理误差**)或者回应者给了错误答案(**回应误差**)，也都属于非抽样误差的例子。最后一点，连问题的措辞都对答案有重大影响。设计抽样调查的人会应用一些统计技巧来设法减低非抽样误差，他们也会用比简单随机样本复杂的**概率样本**，比如**分层样本**。只要你观察一些基本事项就可以对一项抽样调查的质量做相当好的评估，这些事项包括是否用随机样本，样本大小和误差界限、无回应率以及问题的措辞。



第4章 习题

4.1 不在误差界限之内。从最近的一项盖洛普调查得知,68%的美国成年人,赞成在公立学校同时教授进化论及神创论。盖洛普发表新闻时说:

对于这样大小的样本所得的结果来说,我们有95%的信心可以宣称,由抽样和其他随机效应所产生的最大可能误差,是正负3个百分点。

对于调查结果的可能误差来源,举一个例子是没有包含在误差界限内的。

4.2 哪种误差?以下何者是抽样误差的来源,何者是非抽样误差的来源?说明你的答案。

- (a) 受访对象隐瞒曾用过毒品的事实。
- (b) 记录资料时打字错误。
- (c) 通过要求人们寄回印在报纸上的购物优惠券来搜集资料。

4.3 哪种误差?以下每一项都是抽样调查的误差来源。把每一项归类为抽样误差或非抽样误差,并且说明理由。

- (a) 用电话号码簿当作抽样框。
- (b) 打了5次电话都联络不到受访者。
- (c) 访问员在街上找人访问。

4.4 互联网使用者。有一项对互联网使用者的调查,发现男、女使用者的比例是2比1。这个结果有点出人意料,因为之前的调查结果中,男、女比例是9比1。把这篇报道继续读下去,就读到以下信息:

我们送出详细的问卷给网络上超过13 000家机构;共收到1 468份有效回应。根据寇特门先生所言,误差界限是2.8%,置信水平为95%。

- (a) 此调查的回应率是多少?(回应率是指预定的样本中,实际回应的百分比。)
- (b) 用速算法来估计此项调查的误差界限。你的结果和他们声明的



2.8% 接近不接近?

(c) 你觉得这个颇小的误差界限, 是不是此调查结果精确程度的合理估计? 请说明你的答案。

4.5 抽取学生样本。某大学从注册组的全部大学生名单中, 抽出了一个 100 人的 SRS, 来访问有关大学生的问题。如果他们同时选了两个 100 名学生的 SRS, 从这两个样本得到的结果, 应该会有些不同。这种变异属于抽样误差或是非抽样误差的来源? 调查结果当中的误差界限, 有没有把这种误差来源考虑进去?

4.6 铃铃铃没人应。电话访问中常出现的一种无回应是“铃铃铃没人应”。也就是说, 拨电话到一个有效号码, 但是没人接听。意大利国家统计局(Italian National Statistical Institute)检视了 1 月 1 日到复活节, 以及 7 月 1 日到 8 月 31 日这两段期间内, 意大利政府向住户所做的一项调查的无回应状况。所有的电话都是在晚上 7 点到 10 点之间打的, 但是有一段期间有 21.4% 的电话“铃铃铃没人应”, 另一段期间有 41.5% “铃铃铃没人应”。你认为比较高的没人应比例是哪一段时间的? 为什么? 并说明为什么无回应率高时, 样本结果较不可靠。

4.7 他该走还是该留。1998 年 12 月的时候, 美国众议院弹劾了克林顿总统, 这是免除总统职务的第一步。然后美国参议院进行了审判, 并判总统无罪。以下是在众议院采取行动之后, 民意调查中提出的两个问题:

你觉得克林顿总统应该怎么做: 在参议院对被指控的事项抗争? 还是干脆辞职?

你觉得克林顿总统应该怎么做: 继续总统职务并在参议院面对审判? 还是干脆辞职?

对第一个问题的回应是, 有 58% 的人觉得总统应该辞职。但是被问到第二个问题的人当中, 只有 43% 觉得他应该辞职。你觉得第一个问题的措辞, 为什么鼓励了更多人赞成总统辞职?

4.8 修订宪法。你正在为一项有人提出的宪法修正条款, 构思民意调查的问卷题目。你可以问人们是否愿意同意修正案来“改变宪法”或者“增加宪法条款”。这两种说法中, 哪一个会造成高得多的同意



比例?为什么?

4.9 老百姓想要减税吗?在2000年美国总统大选前,总统候选人对于政府的高额节余应如何处理展开辩论。普优研究中心对于美国成年人的随机样本问了两个问题。两个问题当中都说社会保险(Social Security)将“保持不变”。对于节余剩下的部分,有以下两种建议:

多的钱应该用来减税?还是用来资助政府的新计划?

多的钱应该用来减税?还是应该用在教育、环境、医疗、打击犯罪及国防等各项计划上面?

这两个问题中,一个问题的回答是有60%的人赞成减税,另一个问题的回答却只有22%的人赞成减税,哪一个问题把回应的人拉向减税?为什么?

4.10 问卷题目的措辞。讨论底下这些可能的问卷题目。问题清楚吗?措辞有没有倾向某一个答案?

(a) 以下哪一项最能代表你对枪械管制的意见?

1. 政府应该把我们的枪都收走。
2. 我们有权利拥有并携带枪械。

(b) 美国应该建立一个洲际导弹防御系统,因为如此可让我们免于恐怖分子的威胁,并居于世界领导地位。你同不同意?

(c) 有鉴于日渐增长的环境破坏问题,以及已现端倪的资源枯竭现象,你赞不赞成对于回收资源密集产品给予经济上的奖励?

4.11 坏问卷题目。自己写写看坏抽样问卷题目。

(a) 写一个“偏心”的问题,让大部分人的回答会偏向两个答案中的某一个。

(b) 写一个语意混淆不清、很难回答的题目。

4.12 评估一项调查。《华尔街日报》(*The Wall Street Journal*)刊登了一篇有关美国大众对社会保险制度态度的文章,其内容是根据一项抽样调查得来的。调查结果发现,比如说,18—34岁的人当中,有36%不期望在退休时得到社会保险制度的任何补助。新闻报道通常对于抽样调查的描述都很简短。以下是《华尔街日报》对此次调查描述的一部分:

华尔街日报/NBC电视台的民意调查结果,是根据电话



访问了全国各地共2012位成年人所得到的，访问是由赫提民意调查公司在周四至周日之间进行的。

样本是从随机选自美国大陆的520个地理区中选出，从每个地理区取的样本大小，和该区人口成正比。选住户时所用的方法，让每个电话号码，不论有没有列在电话簿上，都有一样的中选机会。

列了很多项对于民意调查该问的问题。《华尔街日报》对这些问题给了什么样的答案？

4.13 估计一项调查。《纽约时报》上一篇有关政党倾向的文章，讨论了一项抽样调查的结果，结果中包括：有44%成人认为民主党“对于如何带领国家进入21世纪比较有想法”。另有37%的人选择共和党；其他人没意见。以下是时报关于“民意调查如何进行”内容的一部分：

最近一次纽约时报/NBC电视台民意调查，用电话访问了美国各地1162位成人，访问时间是11月4日至7日……

电话交换机样本，是用电脑从全国超过42000个有效住宅交换机中随机选出。

在每个选中的交换机内，再把随机数字加上，构成完整的电话号码，如此则不论有没有列在电话号码簿上的电话号码都可以包括在内。在住户之内，亦用随机方法选定一位成人接受访问。

列了很多项关于相信调查结果之前该问的问题。纽约时报对这些问题给了怎样的答案？

4.14 封闭型及开放型问题。问题基本上可分封闭型及开放型两种。封闭型问题会提供一组固定回答让受访者选。开放型问题让受访者用自己的话回答。访问员把回答抄录下来，之后再分类。开放型问题的例子如下：

你觉得青花菜如何？

封闭型问题的例子是：

你觉得青花菜如何？是

- a. 非常喜欢？
- b. 喜欢？
- c. 不喜欢也不讨厌？



d. 有点讨厌?

e. 非常讨厌?

开放型问题和封闭型问题各有什么利弊?

4.15 有没有说实话?许多受访者对于牵涉到不法活动,或者是敏感问题,都不诚实回答。有项研究把一大群白种人,随机等分成三组。每个人都被问到是否曾使用可卡因。第一组人用电话访问,21%说“有”。一组人由访问员到家里访问,25%说“有”。最后一组也是到家里访问,但是回答是填在没有记号的表格上,然后密封在一个信封里。这组里面有28%说曾使用过可卡因。

(a) 你觉得哪一个结果最接近事实?为什么?

(b) 你举两种行为的例子,你觉得用电话访问得到的比例会比实际的低。

4.16 你投票了吗?“当前人口调查”询问美国50 000个住户样本中的成人,有没有在1996年的总统选举中投票,54%说有。事实上,整个成人总体中,只有49%在那次选举中投了票。CPS的结果和真正值的差距,比误差界限大很多,你认为这是为什么?

4.17 在宴会中抽样。某个宴会当中有30位21岁以上的学生,以及20位不到21岁的学生。你从21岁以上的人中随机抽出3位,不到21岁的人中随机抽出2位来访问,问他们对酒的态度。你给了宴会中每个学生同样大的受访机会:这个机会是多大?为什么你的样本不是SRS?

4.18 分层样本。某个俱乐部有30名学生会员及10名教师会员。学生名单如下:

亚伯	费雪	胡柏	罗培兹	瑞曼
卡尔森	葛许	杰米尼兹	米兰达	山多斯
陈	葛理斯沃	琼斯	奈曼	萧奥
大卫	海恩	金	欧布莱恩	汤普森
德明	赫南德兹	克劳兹	普尔	乌慈
依勒斯浩	荷兰	刘	帕特	伐格



教师名单如下:

阿里亚戈	佛南德兹	金	墨尔	韦斯特
贝西可维其	辜普格	赖特曼	维卡瑞欧	杨

俱乐部可以送4位学生和2位教师去参加一个大会。他们决定用随机选择的方式来决定谁去。

- (a) 用表A来选一个包括4位学生及2位教师的分层样本。
- (b) 姓山多斯的学生被选中的机会是多大?姓金的教师被选中的机会是多大?

4.19 分层样本。一所大学有2000位男教师和500位女教师。主管公平雇用(equal opportunity employment)的官员,想要从教师中抽出随机样本,来听取他们的意见。为了对女性教师的意见更加重视,他决定要抽一个包含200位男性及200位女性的分层样本。他有按照英文字母排序的女性教师名单及男性教师名单。

- (a) 说明要怎样编代码和用随机数字选取需要的样本。从表A的列122开始,列出你样本中的头5位女教师及头5位男教师。
- (b) 2000位男教师中任何一位会被你选进样本的机会是多大?500位女教师中任何一位会被选进样本的机会是多大?

4.20 会计师的抽样。会计师在替公司查帐时,会用分层样本来查证诸如应收账款之类的记录。分层的依据是每个项目的金额,通常会金额最高的几项全部放进样本。某家公司提出5000项应收账款。其中有100项的金额超过50000美元;500项在1000—50000之间,其余4400项低于1000元。把这3组不同金额当做层,假如你决定金额最高的层中,每一项都要查证,中等金额的层查证5%,低额的层只查1%。你要从中抽样本的这两层要怎样编代码?用表A,从列115开始从这两层各选出头5笔账。

4.21 抽样有矛盾?例8比较了两个SRS,分别是一所大学的大学生及研究生。大学生样本占大学生总体的比例比较小,只有1/90,而研究生样本则占研究生总体的1/15。然而在大学生中90个取1个的抽样误差,反而比在研究生中15个取1个的还要小。向一个不懂统计的人解释,为什么会这样。



4.22 习题 4.12 中有《华尔街日报》对于一项抽样调查的部分描述,从中看来,样本是分几个阶段抽取的。为什么我们可以这样说?第一阶段无疑是用了分层样本,虽然日报并未明说。说明一下,为什么从全国一大堆地区中抽出 SRS,而不用分层样本,并不是个好主意?

4.23 《科学》(*Science*)期刊中有一篇文章,探讨欧洲和美国两者之间,对于基因改造食物的态度有无不同。这就得做抽样调查了。欧洲的调查在欧洲 17 国中,各抽了一个 1 000 人的样本。以下是部分描述:“欧洲气压计(Eurobarometer)调查是一项多阶段,随机概率(random-probability)的面对面访谈。”

(a) 多阶段是什么意思?

(b) 你可以看到第一阶段是分层抽样。有哪些层?

(c) “随机概率样本”(random-probability sample)是什么意思?

4.24 小学生想要什么?小学生的主要目标是什么?女生和男生有不同的目标吗?城市、郊区和乡村的小学生,目标有不同吗?为了找答案,研究者希望能问四、五、六年级的小学生以下问题:

你在学校最希望能做到什么?

a. 有好成绩。

b. 运动出色。

c. 人缘好。

因为大部分儿童住在人口稠密的城市及郊区,如果取一个 SRS,可能会几乎包含不到乡村儿童。而且从很大一个区域随机抽取儿童,调查的花费很大,最好是先抽学校,再抽儿童。提出一个合适的抽样设计,并说明为什么你的设计适合这项研究。

4.25 系统随机样本(systematic random sample)。“当前人口调查”(例 7)的最后阶段用了系统随机样本。下面这个例子可以说明系统样本是怎么回事。假设我们得在宿舍的 100 个房间中选出 4 间。因为 $100/4=25$,我们可以把 100 间房间的清单,想成是 4 个清单,每个有 25 间房。用表 A 从头 25 间房中随机选一间。样本会包括选出的这间以及这间之后的第 25、第 50 及第 75 间。比如说如果先选中 13 号房,则系统随机样本就包括号码为 13、38、63 及 88 的这几间房。用表 A 从 200 个房间的清单中,选出一个包含 5 间房的系统



随机样本。从列 120 开始。

4.26 系统样本非简单随机样本。例 4.25 描述了一个系统随机样本。和 SRS 一样，系统随机样本也给每个个体同样被选中的机会。说明为什么是这样，并详细解释为什么即便如此，系统样本却并不是 SRS。

4.27 计划一项对学生的抽样调查。学生代表会想要从学生当中抽出随机样本，了解他们最希望对校园生活做哪些改变？学校提供了 3 500 名学生的注册名单当做抽样框。

- (a) 你会怎么选出一个 250 位学生的 SRS？
- (b) 你会怎么选出一个 250 位学生的系统样本？(参考习题 4.25 对系统样本的说明。)
- (c) 学生名单上注明了是住校(2 400 名学生)还是未住校(1 100 名学生)。你会怎样抽取一个包含 200 位住校生及 50 名不住校学生的分层样本？

4.28 从学生当中抽样。你想要知道和你同校的学生对于学校对性骚扰的处理原则有什么看法。你申请到经费，足够访问 500 名学生。

- (a) 严格定义你这项研究的总体是什么。比如说，推迟毕业的学生算不算？
- (b) 说明你的抽样设计。比如说你会不会用一个以性别分层的分层样本？
- (c) 简单讨论一下，你预期会发生什么实际的困难？比如，你要怎样联络你样本里的学生？

4.29 购物中心访谈。第 2 章的例 1 中，描述了在购物中心的访谈。这是方便样本的例子。为什么购物中心访谈中用的样本，不是概率样本？

4.30 同性恋者参军？以下是针对同一个议题的 3 个民意调查问题，以及民意调查结果：

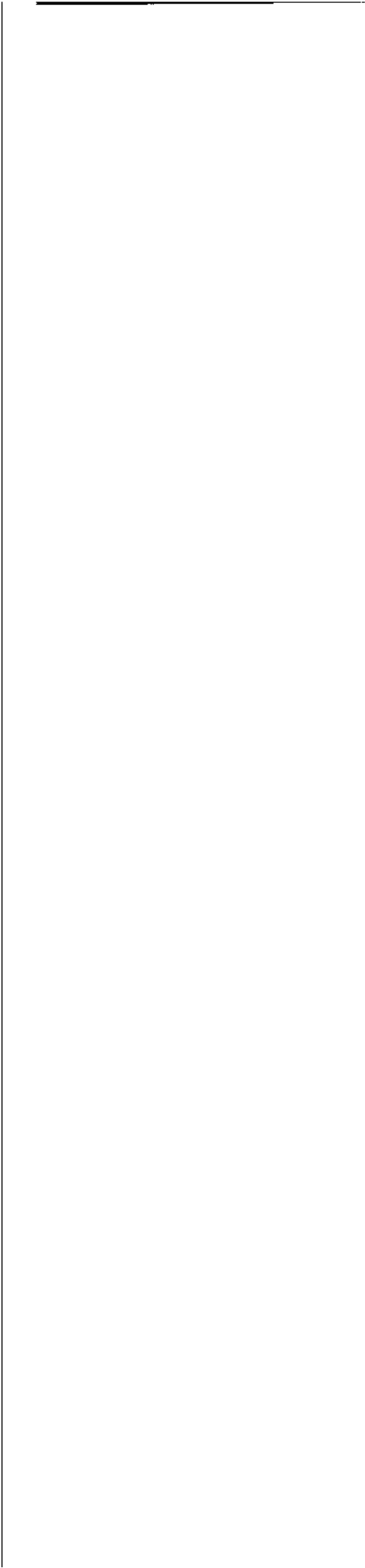
你同不同意，让公开承认自己是同性恋的男性及女性，加入美国军队？结果：47% 很不同意或相当不同意，45% 很同意或相当同意。



你认为应不应该禁止男同性恋者和女同性恋者参军?结果: 37% 说应该禁止; 57% 说不应该。

克林顿总统是否应该改变军队政策, 容许同性恋者加入军队? 结果 53% 说“否”, 35% 说“是”。

用这个例子来说明, 要用对民意调查的回应来了解大众的意见, 是很困难的。



第 5 章

实验面面观

三个例子

有一项关于网上学习的乐观报道，报告了在佛罗里达州劳德代尔的诺瓦东南大学(Nova Southeastern University)执行的一项研究。撰写研究结果的人声称，学生在网上(online)学习大学部的课，和在教室里学习的学生“学得一样好”。如果把教室里的课用网站取代，可以替大学省很多钱，所以照这项研究结果看来，我们应该全部上网。

胃溃疡似乎是一种现代病。“胃冷冻”(gastric freezing)是治疗胃溃疡的一种聪明疗法。病人先吞下一个连接着管子且放了气的气球；然后就把一种经过冷冻的溶液打入气球当中，总共打一小时。这个疗



法的想法是这样的：使胃凉下来可以减少胃酸的分泌，因此可以减轻溃疡症状。一篇刊登在《美国医学会期刊》(*Journal of the American Medical Association*)的实验报告指出，胃冷冻的确缓解了溃疡的痛苦。

美国政府应不应该为低收入户提供日间托儿照护呢？如果这项服务能帮这些儿童日后受更多教育，并且有好的工作，则政府也可以因为少付出福利金以及增加税收而省钱，因此连那些精打细算的纳税人都可能支持日间照护计划。卡罗来纳启蒙计划曾从1972年开始，持续追踪观察一群儿童。结果显示，好的日间照护计划，对儿童以后的就学及就业，有长足的影响。

刚才说到三个议题，以及三项针对那些议题做的研究，研究结果都对议题有了某种结论。而事实上，第一项研究是观测研究，其对于大学的网上学习所做结论是靠不住的。胃冷冻实验现在看来只是误导的结果，因为大部分溃疡是由细菌造成的，把胃弄凉不会有什么作用。然而启蒙计划的结果，已经差不多是对于日间照护的长远效果所可以找到的最好证据了，是什么因素使得有些研究(尤其是实验)令人信服呢？为什么其他的研究我们可以置之不理呢？本章会告诉我们重点在哪里。

谈谈实验

观测研究是被动的数据搜集方式。我们只观察、记录或度量，但是不干扰。而实验却能主动产生数据。做实验的人会主动介入，他会把某项处理加诸于受试对象，来观察受试对象有何反应。所有的实验以及许多观测研究，都是想要知道一个变量对另一个变量有何影响。

以下是我们用来分辨哪个变量是影响者，哪个变量是被影响者。



实验用语

反应变量是指用来度量研究结果的变量。

解释变量是我们认为可以解释或造成反应变量变化的变量。

实验中所研究的个体，通常称为**受试对象**。

处理是任何加诸于受试对象的特定实验条件。若实验当中有数个解释变量，则处理就是指每个变量都设定一个特定值后的组合。

例 1 上网学习及日间照护的效果

大学生是诺瓦东南大学研究中的受试对象，而解释变量是学习环境(在教室或是在网上)；反应变量是修完课后，学生的考试成绩。

启蒙计划是一项实验，其中的受试对象是 111 个人，这些人在 1972 年时仍是婴儿，出生在北卡罗来纳州教堂山(Chapel Hill)一些低收入户黑人家庭中，且身体健康。所有这些婴儿都得到社会工作者的帮助以及营养补充；其中并随机选出一半人，给予密集学前教育。这项实验比较了这两项处理。解释变量只是“学前教育，有或者没有”。反应变量则有很多，且记录的时间超过 20 年，其中包括学力测验成绩、是否上大学以及就业情况。

很多地方把解释变量叫做独立变量(independent variable, 或称自变量)，反应变量叫做因变量(dependent variable)。用这样的名称是因为反应变量是跟着解释变量变的。我不大喜欢以前用的这些名词，一部分原因是“独立”这个词在统计里面另有很不一样的意思。

怎么样做坏实验

上网修课的学生，是不是和在传统教室修同样课的学生学得一样好？要知道答案最好的方法，是指定一些学生到教室上课，其他学生上网。这样就是实验。诺瓦东南大学的研究不是实验，因为它没有对



学生受试对象加诸任何处理，而由学生自行选择要在教室上课还是上网。研究中只度量了他们的学习成果。选择上网的学生，原本就和选择到教室上课的学生很不一样。比如说，在修课之前对于课程相关内容的测试当中，上网学生的平均分数是 40.70，而选择到教室上课的学生平均分数是 27.64，在上网学生的水平原本已经大大超前的情况下，很难比较出来教室学习及上网学习的优劣。上网学习以及教室学习的效果，已经和一些潜藏在背景里的因素，无可救药的混杂在一起了。图 5.1 表示了这种混在一起的影响。

• 交叉

潜在变量是对研究中其他变量间的关系有重要影响，却并未被列为解释变数的变量。

当两个变量对反应变数的影响混在一起而无法区分时，我们称这两个变量是**交叉**的。交叉的变量可以是解释变量，也可以是潜在变量。

在诺瓦东南大学的研究中，学生原来的程度(潜在变量)就和解释变量交叉在一起了。研究报告说两组学生在期末考试中考得一样好。我们没法判断，上网学习那组学生的表现，有多大一部分可以归因于他们原来的程度。一组本来程度就很好的学生，和原来比他们差的在教室学习的学生，修完课后比起来，表现却一样，这样的结果，好像不能算是网络课程有神奇效果的证据。以下是另一个例子，在这个例子当中谈到，通过第二项实验，把第一项实验中的交叉状况给理顺了。

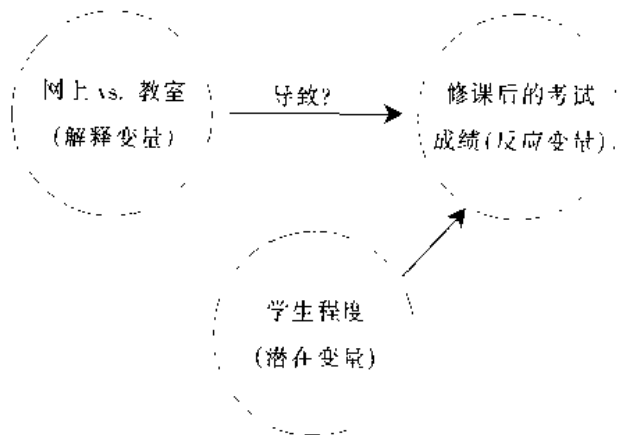


图 5.1 诺瓦东南大学研究中的交叉状况。不同教学环境(解释变量)的影响，和学生程度(潜在变量)对结果的影响，没有办法分辨得出来



例 2 胃冷冻失败了

实际在病人身上研究医疗效果的实验，叫做**临床试验**(clinical trial)。使得胃冷冻变成胃溃疡的一般疗法的临床试验，其架构是一种“单轨”设计：



接受治疗后病人的确表示比较不痛了，但是我们不能据此宣称，胃冷冻使得疼痛减轻。

这有可能只是**安慰剂效应**(placebo effect)。**安慰剂**是一种假的治疗，没有实质效用。许多病人对任何治疗都有正面反应，即使只是安慰剂。这种对假治疗的反应，就称为安慰剂效应。安慰剂效应可能是一种心理作用，起因于对医生有信心以及预期病会治愈。也许这也只是一种说法，用来涵盖许多病人，在无明显原因下病况却改善的事实。实验使用了单轨设计，代表安慰剂效应会和胃冷冻可能会有的任何效应交叉在一起。

数年之后做了另一项临床试验，把溃疡病人先分成两组。一组就像前次试验一样，接受胃冷冻治疗；另一组接受的是安慰剂治疗，也就是打入汽球的溶液温度和体温一样，而不是经过冷冻的。结果是：处理组的 82 位病人中，有 34% 病情改善，但是安慰剂组的 78 位病人中，也有 38% 有改善。这项实验和其他妥善设计的实验，显示出胃冷冻的效应，不过是和安慰剂差不多罢了，于是从此医生不再使用这种方法。

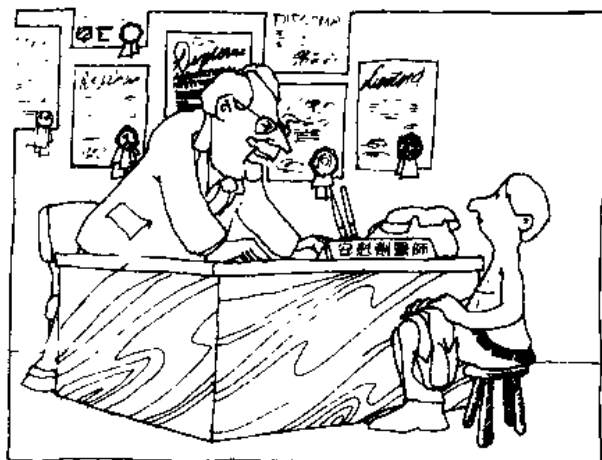
观测研究和单轨实验都常常因为和潜在变量的交叉问题，而产生没有用的数据。如果只能观察，交叉现象很难避免。做实验，情况就好得多，这从刚才说到的第二次胃冷冻实验就可以看出来。这次实验当中，多包括了一组只接受了安慰剂的受试对象，这样让我们可以比较，正在检验的这项治疗，效果是否比安慰剂要好；若结果为是的话，则它的效果应不止是安慰剂效应而已。有效的医疗会通过这项安慰剂测试，胃冷冻则失败了。



随机化比较实验

设计实验的第一个目标,是要确定实验可以显示解释变量对于反应变量的影响。单轨实验常常因为交叉而达不到这个目标。补救的方法是同时比较两个或多个处理。

接下来就是一项新医疗法通过直接比较而通过安慰剂测试的一个例子。



“有件事我想讲得一清二楚,史密斯先生。我给你开的药,会治好你的疲惫感。”

例3 地中海贫血症

地中海贫血(sickle cell anemia),是一种遗传性的红血球异常,在美国得这种病的大多是黑人。它能导致剧痛以及许多并发症。美国国家健康研究所(National Institutes of Health)执行临床试验,用一种叫“羟基脲”(hydroxyurea)的药来治疗地中海贫血。受试对象为299个成年病人,这些人在过去一年当中,都因为地中海贫血而至少有过三次剧痛的发作。

如果光是把羟基脲给所有299个受试对象服用,就会把药效和安慰剂效用及其他潜在变量的效应(例如自知是实验的受试对象所产生的效应)全部混杂在一起。所以只有一半的受试对象服用羟基脲,而另一半服用的是看起来和尝起来都像羟基脲的安慰剂。除了药的内容以外,所有受试对象的治疗过程完全一样(比如说,检查时间的安排都一样)。因此潜在变量会对两组产生同样的影响,对两组的平均回应不会造成差异。

两组受试对象在服药之前,应该在各方面条件都要相近。就跟抽样一样,在我们选择哪些受试对象服用羟基脲时,要避免偏差的最好方法就是,避开人为选择,完全随机决定。我们从所有受试对象中选出大小为152的SRS组成羟基脲组,剩下的147人就组成安慰剂组了。图5.2有这个实验设计的大略描述:

实验比预订时程提早结束,因为羟基脲组的剧痛发作次数比安慰剂组少得多。这已经是足以令人信服的证据,证实羟基脲是地中海贫血的有效疗法,对身受这种严重疾病之苦的人来说,这真是好消息。

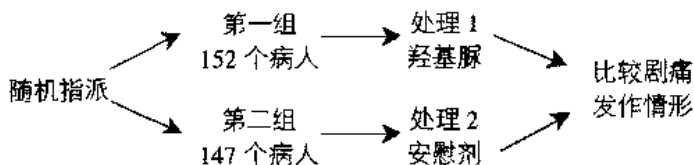


图 5.2 为了将羟基脲对地中海贫血的治疗效果和安慰剂做比较, 所执行的随机化比较实验的设计

图 5.2 说明了最简单的**随机化比较实验**(randomized comparative experiment), 实验只比较两种处理。图里描述了实验设计中的重要信息: 随机指派, 一个处理分配一组人, 每组人数(通常最好是让各组人数接近), 每一组分配到哪种处理以及我们比较的是什么反应变量。你已经知道怎么用随机指派方式分组了: 先把 299 个受试对象编上代码 001—299, 然后从随机数字表(表 A)读出 3 个一组的数字, 直到选出 152 位受试对象来编入第 1 组。剩下的 147 位受试对象就是第 2 组。

例 3 中的安慰剂组叫做**控制组**(control group), 因为通过对处理组和控制组的比较, 使我们能够控制潜在变量的影响。控制组不一定是接受像安慰剂那样的假治疗, 临床试验常常会把新的治疗方法和已经在使用的方法进行比较, 而不是和安慰剂比。随机指派到现有疗法的病人, 就构成控制组。如果要比较的处理超过两个, 我们可以将所有受试对象随机指派到不同组去, 组数和处理数相同。以下是分成 3 组的一个例子。

例 4 节约能源

很多公用事业公司都有鼓励顾客节约能源的方案。有一家电力公司考虑在住宅装一种电表, 电表能够显示: 如果当时的用电量持续整个月, 花费会是多少。这种电表会减低用电量吗? 还是有什么更省钱的方法也有同样的功效? 这家公司决定要设计一个实验来测试。

有一种比较省钱的办法是: 给顾客一张图表以及如何监测用电量的信息。这个实验对这两种方法(电表、图表)及控制组进行比较。控制组的顾客会得到有关节约能源的信息, 但不会得到监测用电量的任何帮助。反应变量是一整年的用电量。公



司在同一个城市中找到 60 个愿意参加实验的单一家庭住宅，所以 3 个处理中的每一个，都各有 20 个被随机指派的住宅。图 5.3 说明了设计概要。

为了执行随机指派，我们将 60 户住宅从 01 到 60 编代码，然后从表 A 中选出含 20 户住宅的 SRS 来接受电表，继续从表 A 中选出 20 户来接受图表，剩下 20 户就当控制组。

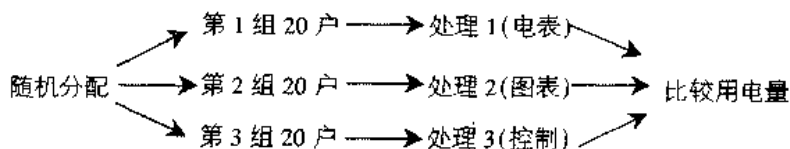


图 5.3 住户节约能源计划三个方案的随机化比较实验设计

实验设计的逻辑

随机化比较实验是统计学里面最重要的概念之一。它的设计是要让我们能够得到明确因果关系的结论。我们先来弄清楚随机化比较实验的逻辑：

- 用随机化的方法将受试对象分组，所分出的各组在实施处理之前，应该各方面都类似。
- 用“比较”的设计以确保：除了实验上的处理外，其他所有因素都会同样作用在所有的组上。
- 因此，反应变量的差异必是处理的效用所致。

我们用随机方法选组，以避免人为指派时可能发生的系统性偏差。例如，在地中海贫血的研究中，医师有可能下意识就把最严重的病人指派到羟基脲组，指望这个正在试验的药能对他们有帮助。那样就会使实验有偏差，不利于羟基脲。从受试对象中取 SRS 来当做第 1 组，会使得每个人被选入第 1 组或第 2 组的机会相等。我们可以预期两组在各方面都接近，例如：年龄、病情严重程度、抽不抽烟等等。



举例来说，随机性通常会使两组中的吸烟人数差不多，即使我们并不知道哪些受试对象吸烟。

要是告诉你，医学研究者对于随机化比较实验接受得很慢，应该不会让你惊讶，因为许多医师认为一项新疗法对病人是否有效，他们“只要看看”就知道。但事实才不是这样。有很多医疗方法(例如胃冷冻)只经过单轨实验后就普遍使用，但是后来有人起疑，进行了随机化比较实验后，却发现其效用充其量不过是安慰剂罢了，这种例子已经不胜枚举。医学文献里可以找到，经过适当的比较实验研究过的疗法，以及只经过“历史对照组”(historical control)实验的疗法。用历史对照组做的研究不是把新疗法的结果和控制组比，而是和过去类似的病人在治疗后的效果作比较。纳入研究的 56 种新疗法当中，用历史对照组来比较时，有 44 种疗法显示出有效。然而在经过使用合适的随机化比较实验后，只有 10 种通过安慰剂测试。即使只跟过去的病人比，医师的判断仍过于乐观。目前来说，法律已有规定，任何新药必须用随机化比较实验来证明其安全性及有效性。但是对于其他医疗方法，比如手术，就没有这项规定。你可以指望新药一定比安慰剂好，但新的手术概念就未必了，这情况就和以前的胃冷冻一样。

对于随机实验有一件重要的事必须注意。和随机样本一样，随机实验照样要受机遇法则(law of chance)的“管辖”。就像抽一个选民的 SRS 时，有可能运气不好，抽到的几乎都是共和党员一样，随机指派受试对象时，也可能运气不好，把抽烟的人几乎全放在同一组。我们知道如果抽很大的随机样本，样本的组成和总体近似的机会就很大。同样的道理，如果我们用很多受试对象，则利用随机指派方式分的组，也就有可能有类似实际情况的组成。受试对象较多，表示处理组(treatment group)的机遇变异会比较小，因此实验结果的机遇变异也会比较小。“用足够多的受试对象”和“同时比较数个处理”以及“随机化”，同为“统计实验设计”(statistical design of experiment)的基本原则。

• 实验设计的原则

统计实验设计的基本原则如下：

1. 要控制潜在变量对反应的影响，最简单的方法是同时比较至少 2 个处理。
2. 随机化：用非人为的随机方法来指派受试对象到不同的处理组。
3. 每一组的受试对象要足够多，以减低结果中的机遇变异。



统计显著性

因为机遇变异的存在，让我们应该更仔细看看随机化比较实验的逻辑。我们不能说，只要羟基脲组和控制组的患者剧痛发作的平均次数有差别，就一定是因为药的效用所导致。就算两组用完全一样的处理，结果也是会有差别的，因为受试对象永远会有个别差异。

即使随机化可以消除组与组之间的系统差异，机遇差异还是存在。我们应该要求反应变量间的差异要大，使得差异不会仅因机遇变异就发生。

• 统计显著性

我们观察到的效果如果大到某种程度，光靠机遇产生这种结果的概率很小时，我们就称此结果有**统计显著性**(statistical significance)。

羟基脲组和控制组之间，剧痛发作的平均次数差别已经有“高度统计显著性”(highly statistically significance)。这个意思是说，这么大的差别几乎不可能全靠机遇产生。我们的确有很强的证据，证明羟基脲对地中海贫血患者的帮助胜过了安慰剂。在很多不同研究领域的调查报告当中，你都常常会看到“有统计显著性”这样的用语。这就是告诉你，对于想要证明的效用，调查者找到好的证据了。

当然实验的实际结果，比起得到统计显著性的认证这件事，要更加重要。在地中海贫血的实验当中，处理组一年里剧痛发作的平均次数是2.5，而控制组是4.5。有这样大的差距，对病患来说是很重要的结果。如果差别只是2.5对应2.8，则即便是有统计显著性，这个结果也无足轻重。

只能观测的时候怎么办

按时去教堂会使人寿命较长吗？医师在治疗心脏病时，对女性有歧视吗？一边开车一边打移动电话，会增加出车祸的概率吗？这些都是



因果问题(cause-and-effect question), 所以应该用我们喜爱的招式: 随机化比较实验。可是很抱歉。我们不能随机指派某些人去教堂, 某些人不去, 因为去不去参加宗教活动是个人信仰问题。我们也不能用随机数字表, 来随机指定心脏病患者是男性或女性。而要求驾驶员一边开车一边用移动电话, 是我们不愿意做的事情, 因为边打电话边开车可能比较危险。

对于以上这些问题, 以及许多其他因果问题, 我们能得到的最好数据, 是从观测研究得来的。我们知道观测是次于实验的第二选择, 而所得结果比实验结果弱得多, 但是好的观测研究可绝不是一无用处。那怎么样的观测研究才算好的呢?

首先, 好的研究不管是不是实验, 都一定要做比较。我们可以从固定做礼拜的人和没有固定做礼拜的人当中, 各自抽随机样本出来比较。也可以比较医师如何治疗男病人和女病人。也许我们可以比较同样一个人, 在开车时移动电话和不打电话时的情况。我们常常可以通过同时运用比较和适配(matching), 而创造出控制组。为了想要知道怀孕期间服用止痛药的影响, 我们比较了服用止痛药和未服止痛药的女性。我们从未服药的许许多多女性中, 选出一些女性, 在年龄、教育背景、子女数以及其他潜在变量的各方面, 都和服止痛药那组女性很接近。这样我们就有两组女性, 在所有这些潜在变量的各方面都近似, 所以这些潜在变量就应该不会影响我们对于两组女性的比较结果。

比较并不能消除交叉。按时去教堂或犹太会堂或清真寺的人, 比不去的人更会照顾自己。他们当中较少人抽烟, 较多人运动, 也比较少人超重。适配可以缩小某些差距, 但不是所有差距。如果把去教堂的人和不去教堂的人去世时的年龄直接做比较, 就会把宗教的影响和良好生活习惯的影响交叉在一起。所以好的比较研究, 必须能够**度量并且调整交叉变量**。如果我们度量体重、抽烟习惯、运动习惯, 就可以用统计技巧来减少这些变量对寿命的影响, 而只剩下(我们希望如此)宗教的影响。

例 5 通过宗教来长寿

针对按时参与宗教礼拜的效果所做的较好研究之一, 搜集了 3 617 位成人的随



机样本。随机样本是好的开始。然后研究者除了解释变量(宗教活动)和反应变量(寿命长短)之外,还度量了许多其他变量。新闻报道说:

去教堂的人有较大比例不抽烟、常运动以及体重适中。不过即使把健康习惯列入考虑,没有按时做礼拜的人,还是比按时做礼拜的人,多25%的死亡机会。

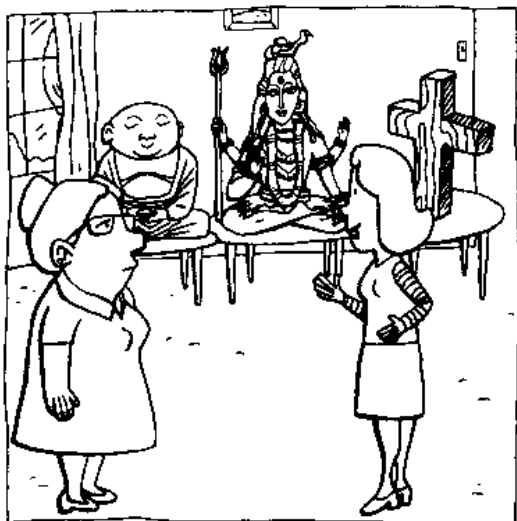
“列入考虑”的意思是说,最后得到的结果,曾经根据两组的差异做调整。这些调整减低了宗教的优越程度,但是仍然留下了相当大的优势。

例6 治疗心脏病有性别歧视?

医师对于女性心脏病患者,比较不常像对同样症状的男性患者那样,给予积极治疗。这是不是代表医师有性别歧视?未必见得。女性的心脏问题,通常比男性晚发生,因此女性心脏病患者年纪比较大,也通常还有别的健康问题。这也许可以解释,为何医师治疗她们时更小心。

这个情况需要做一次比较研究,并用统计方法来调整交叉变量的影响。类似的研究已有许多人做过,但结果却相互矛盾。有的结果以医师的话来说是:“当男患者和女患者除性别以外其他方面都相同时,治疗方式是很接近的。”其他的结果却指出,即使已根据男患者和女患者的差距调整之后,女患者接受的治疗还是比较少。

从例6可以看出,统计调整是颇微妙的。随机选择可以制造出对于所有已知或未知变量都接近的组。而适配和调整,对研究中没有考虑要度量的变量,并不产生效用。即使你相信研究者什么都考虑到了,还是要对统计调整稍微存疑。要决定调整哪些变量,有很多作弊空间。而且“经过调整”的结论,实在是等于在说:



“根据统计的结果，有宗教信仰的人较长寿，所以我每星期7天，每天各进行一种宗教仪式，确保我受庇荫。”

如果女性心脏病患者年纪轻些，身体健康些，而男性心脏病患者老一点，健康差一点，那么两种性别的患者会得到差不多的治疗。

也许最好就只能做到这样，而且还要感谢有统计才能达到这种智慧。不过，这可真让我们想念起好实验做出来的清清楚楚的结果了。

网络寻奇

你可以在《美国医学会期刊》的网站 Jama.ama-assn.org 以及《新英格兰医学期刊》(*New England Journal of Medicine*) 的网站 www.nejm.org 中找到最新的医学研究资料。里面许多文章都描述了随机化比较实验，而用到统计显著性这样的语言的更多。



本章重点摘要

统计研究常常试图提出证据证明,当改变某个变量(**解释变量**)的时候,会使另一个变量(**反应变量**)产生变化。在实验当中,我们会自己设定解释变量,而不是只观察它们。观测研究和只实施一种处理的单轨实验,因为和**潜在变量**有交叉,不可能分得出处理的效果到底是什么,所以通常无法生产出有用的数据。补救方法是利用**随机化比较实验**,比较两个或更多个处理,利用机遇决定哪些受试对象接受哪个处理,并且用足够多的受试对象,使得机遇产生的影响变小。比较两个或多个处理,可以**控制**诸如像**安慰剂效应**等的潜在变量,因为潜在变量对每个处理组都同样有作用。

不同的处理所产生的效应差距,若大到几乎没有可能仅因为机遇而产生时,叫做有**统计显著性**。从随机化比较实验所得到具统计显著性的结果,是改变解释变量会**导致**反应变量改变的最好证据。对于因果问题做的观测研究,如果能**比较相似的组**,并且尽量多度量潜在变量来做**统计调整**的话,结果会比较可信。对于回答因果问题的研究来说,观测研究是远落后于实验的第二名。



第5章 习题

5.1 治疗乳癌。对于早期发现的乳癌，怎样治疗比较好？曾经有段时间，最普遍的疗法就是切除乳房。现在则通常只拿掉肿瘤及附近的淋巴结，然后做放射线治疗。为了研究这两种疗法的效果是否有差别，一个医疗小组检查了25家大医院的病历，并且比较使用两种疗法进行治疗的女性在手术之后的存活时间。

(a) 解释变量和反应变量是什么？

(b) 仔细说明为何这项研究不是实验。

(c) 说明为什么交叉会阻碍这项研究找出哪个疗法较有效的真相。

(目前的治疗方法，事实上是在经过一项大规模的随机化比较实验之后，才受到推荐。)

5.2 教导阅读。某位教育工作者想要比较，教导如何阅读的电脑软件和传统的阅读课程，何者较有效。他先测验了一班四年级学生的阅读能力，然后把他们分成两组：一组用电脑学习，另一组参加传统课程。一年之后，他再对学生做测验，并比较两组学生所增加的阅读能力。这是不是实验？理由是什么？解释变量和反应变量是哪些？

5.3 自己哺乳。女性杂志里的一篇文章说，自己哺乳的母亲和用奶瓶喂奶的母亲比起来，对自己的婴儿会觉得更亲近，更容易接受。作者下结论说，自己哺乳对于母亲对婴儿的态度有正面效果。但是要自己哺乳还是用奶瓶喂，是女性自己决定的。说明为何这项事实会使得任何有关因果的结论不可靠。说明时把潜在变量及交叉这些字汇用进去，并且画一个像图 5.1 的图来说明。

5.4 职业训练有用吗？美国某州为制造业的失业工人设置了职业训练计划。5年之后，该州政府评估该项计划的成效。有评论员说，因为该州制造业工人的失业率在计划开始时是6%，而5年之后是10%，所以该计划没有用。

说明为什么失业率提高不见得代表职训计划失败。有哪些潜在变量对失业率的影响，可能会和职训计划的效果发生交叉的，请提出来。画一个类似图 5.1 的图来说明你的解释。



5.5 阿司匹林和心脏病。阿司匹林可以防止心脏病发作吗?“医师健康研究”(The Physicians' Health Study)这项有 22 000 位男医师加入的大规模医学实验,就试图回答这个问题,大约 11 000 个医师的一组每隔一天吃一颗阿司匹林,其他人吃安慰剂。数年之后,该研究发现,阿司匹林组的心脏病发作次数,比安慰剂组的少得多。

- (a) 指出实验中的受试对象,解释变量是什么以及该变量可能的值,还有反应变量是什么。
- (b) 用图表来描述“医师健康研究”的设计。(要描述一项实验的设计时,必须把各处理组的大小以及反应变量都标示出来。图 5.2 和图 5.3 的图可当作范本。)

5.6 学习市场机制。你的经济学教授不大确定,对于让学生上网玩市场游戏,能不能帮助学生了解市场价格是怎么决定的。你建议了一项实验:让一些学生上网玩游戏,其他的则在讨论课里讨论市场机制。这门课的正课有两班,一班早上 8:30 上课,另一班下午 2:30 上课。每一班各分成 10 个讨论小组,学生都已分好组了。为了方便起见,同一个讨论小组的学生,应该都参加同一个计划。教授说:“就让 8:30 那班学生都在讨论课时上网,2:30 那班学生都在讨论课时讨论好了。”为什么这个主意不大高明?

5.7 宣传的效用。1940 年时,有位心理学家进行了一项实验,来研究宣传是否会影响社会大众对外国政府的态度。他对一群美国学生做了一项对德国政府态度的测验。在这些学生读了几个月关于德国的宣传资料后,再测验他们,看看态度是否有改变。

不幸的是,就在实验进行期间,德国攻击并且征服了法国。仔细说明,为什么交叉使得要决定宣传资料是否影响态度这件事,变得不可能。用一个像图 5.1 一样的图来辅助说明。

5.8 再论学习市场机制。

- (a) 习题 5.6 当中提到一个比较两种学习市场机制方法的设计,请约略描述一个会优于那项设计的实验设计。把 20 个讨论小组当做 20 个个体。你觉得应该用什么当做反应变量?(描述一项实验设计时,一定要明白表示出各个处理组的大小,以及反应变量是什么。可以用图 5.2 和图 5.3 当中的图当作范本。)
- (b) 用表 A,从列 116 开始,执行你的设计中必须具备的随机化部分。



5.9 上网学习。接在例1之后的讨论指出,诺瓦东南大学的研究没办法对上网学习和教室学习何者效果较佳,做出什么有效结论,因为选择上网学习的学生原本程度就较好。说明一下怎样可以设计一个比较好的实验,以对这个主题取得有用的信息。

5.10 抗氧化剂可以预防癌症吗?吃很多水果和蔬菜的人,比起吃得少的人,结肠癌的罹患率较低。水果和蔬菜富含像维生素A、C及E等的“抗氧化剂”。服用抗氧化剂能预防结肠癌吗?一项临床试验用了864位属于结肠癌危险群的人来研究这个问题。受试对象被分成四组:一组每天服用 β 胡萝卜素,一组每天服用维生素C及E,一组每天服用 β 胡萝卜素、维生素C及E,一组每天服用安慰剂。经过4年之后,研究者非常意外的发现,四组的结肠癌罹患率,居然没有具统计显著性的差别。

- (a) 这项实验当中的潜在变量和反应变量是什么?
- (b) 描述这项实验设计的概要。(用图5.2和图5.3的图当范本。)
- (c) 给864位受试对象编代码,并且从表A的列118开始,选出 β 胡萝卜素组的头5个受试对象。
- (d) 用“没有具统计显著性的差别”来说明研究结果,究竟是什么意思?
- (e) 想想看有哪些潜在变量可能可以解释,为什么吃很多水果蔬菜的人有较低的结肠癌罹患率。这项实验结果显示,可能是这些潜在变量,而不是抗氧化剂,才是我们观察到的吃水果蔬菜的好处的背后真正原因。

5.11 节约能源。例4里面谈到一项实验用来了解是否给一些住户提供电表或者图表,能够减少电的用量。电力公司一位主管反对在实验中加入一个控制组。他的理由是:“只比较去年(那时还没提供电表或图表)和今年同期的用电量,花费会比较少。如果今年住户用电量较少,就代表电表或图表有用。”清楚说明为什么这个设计要比例4里的设计差。

5.12 改进芝加哥的学校。美国国家科学基金会(NSF, National Science Foundation)出钱资助一项叫“体系行动”(systemic initiatives)的计划,协助各城市改革公立教育系统,以期改善学生的学习状况。这计划有效吗?芝加哥的“行动”,重点在改进高中数学教学。计划进



行两年后，该城 60 所中学当中，有 51 所的学生在—项教学标准测验的平均成绩比以前提高了。NSF 的领导阶层说，这就是芝加哥的计划成功的证明。评论员说，这样的结果并不能用来对“体系行动”的效果做任何结论。解释为什么评论员是对的。有哪些潜在变量可能可以用来解释，为什么过了两年平均成绩会提高？

5.13 封装食物。某家食品制造商利用封装衬条，在袋子装满后用—种热钳把顶部封口，顾客要打开包装时得把封住的部分拉开。而热钳的温度，对于拉开封口要用多大的力有些什么影响？要回答这个问题，工程师用了 20 对封装衬条。他们用 250°F、275°F、300°F 和 325°F 这几种温度，各封合了五对封装衬条。然后度量拉开每个封口，要用多少力量。

- (a) 在此实验中研究的个体不是人，那是什么呢？
- (b) 解释变量是什么，有些什么可能值？
- (c) 反应变量是什么？

5.14 减少医疗支出。要是健康保险要求投保者负担部分的医疗支出，是不是大家会把医疗支出减低呢？—项关于这个主题的实验，想要了解健康保险支付的医疗费用比例，对于人们的看病次数或者他们的健康，会不会有影响。此实验中的处理，是 4 种健康保险计划。医疗费用超出某个上限时，4 种计划都给付所有超出上限的部分。在上限之下，则 4 种计划分别给付支出的 100%、75%、50% 及 0%。

- (a) 描述适合这项研究的随机化比较实验的设计概要。
- (b) 简短讨论一下在这样的实验中，可能产生的实际问题和伦理问题。

5.15 封装食物。用图来描述为习题 5.13 中的封装食物实验，所设计的随机化比较实验。从表 A 的列 120 开始，执行设计中的随机化部分。

5.16 如何处理酒后驾驶。一旦有人被裁定为酒后驾驶，由法庭判定的处理方式或处罚的目的之一，是避免同样的情况再发生。向法院建议 3 种可以用的处理方式。然后大致描述—项可以比较 3 种处理的有效性的实验设计。别忘了列出你要度量哪些反应变量。



5.17 统计显著性 某项随机化比较实验的目的，是要检视如果让健康男性增加钙的摄取，是不是能降低血压。受试对象在 12 周期内，有的摄取钙，有的用安慰剂。研究者做了结论：“增加钙那组的血压显著低于 (significantly lower) 安慰剂那组。”结论中的“显著”，指的是有统计显著性。根据这项实验的内容，来说明统计显著性的意思，请当作你是在向一位不懂统计的医师解释。

5.18 统计显著性。某所大学负责奖助学金的单位，问了一些学生关于他们的工作情况和收入的问题。结果报告中说：“以学期当中的收入来说，不同的性别之间有具统计显著性的差异，平均来说男生收入较多。在白人和黑人学生之间，就没有具统计显著性的差异。”用日常用语来说明“有具统计显著性的差异”和“没有具统计显著性的差异”是什么意思。

5.19 钙和血压。有些医学研究者认为，增加钙的摄取可能可以降低血压。你找到 40 个有高血压的男性，愿意当受试对象。

- (a) 大致描述一个合适的实验设计，不要忘记把安慰剂效应考虑进去。
- (b) 受试对象名单如下。利用表 A，从列 119 开始，来执行设计中的随机化部分。把你要给予钙的受试对象名单列出。

阿洛玛	丹门	韩	梁	洛森
阿斯海洛	德尔	霍华德	马度纳多	所罗门
班奈特	爱德华	鲁斯卡	马兹敦	汤普更斯
毕卡里斯	法鲁克	股拉尼	蒙德西	汤森德
陈	法兰西安纳	杰姆斯	欧布莱恩	涂洛克
克里蒙特	乔治	卡普兰	奥格	恩德渥
克伦斯顿	格林	克鲁雪夫	欧洛斯寇	维拉斯
寇蒂斯	桂伦	洛里斯	罗缀格	张

5.20 治疗前列腺疾病。一项大型研究用了加拿大全国医疗系统的记录，来比较前列腺疾病的两种治疗方法。两种疗法之一是传统手术治疗，另一种是不需要手术的新方法。记录中有许多病人的资料，这些病人的医师有些选择手术治疗，有些选择新疗法。研究显示，用新



方法治疗的病人，在8年之内的死亡率，显著比手术组要高。

(a) 再仔细些研究资料之后发觉，上面的结论是错误的。用新方法治疗的病人中多出来的死亡个案，可以用潜在变量解释。你认为怎样的潜在变量可能会和医师是否选择手术治疗产生交叉？

(b) 有300位前列腺病患愿意当受试对象，进行实验来比较这两个疗法的。用图解来表示出随机化比较实验的设计概要。

5.21 祷告和冥想 你在一份杂志里读到“像冥想或者祷告这类不具体的治疗方式，已经由有控制的科学研究(controlled scientific studies)证实，对于诸如高血压、失眠、溃疡及哮喘等疾病有效。”用简单的语言解释，文章中的“有控制的科学研究”是什么意思，以及为什么这样的研究可能会做出冥想和祷告对某些疾病有效的结论。

5.22 运动及心脏病 规律的运动习惯，能减低心脏病发作的风险吗？以下有两个研究这问题的方法。详细说明，为什么第二种设计可以产生比较有价值的信息。

1. 一位研究者找到2000位40岁以上，有规律运动，而且未曾有过心脏病发作的男性。她给其中每一位适配一位各方面条件类似，可是没有规律运动的男士，然后追踪两组人共5年。
2. 另一位研究者找到4000位40岁以上，未曾有过心脏病发作，且愿意参与研究的男性。她指派其中2000位参加有人指导的规律运动计划，其他2000位继续原来的生活习惯。研究者追踪两组共5年。

5.23 麻醉药的安全性。用不同麻醉药的手术病人，有不同的死亡率。有一项观测研究得出4种麻醉药的死亡率如下：

麻醉药	氟烷	喷妥撒	环丙烷	乙醚
死亡率	1.7%	1.7%	3.4%	1.9%

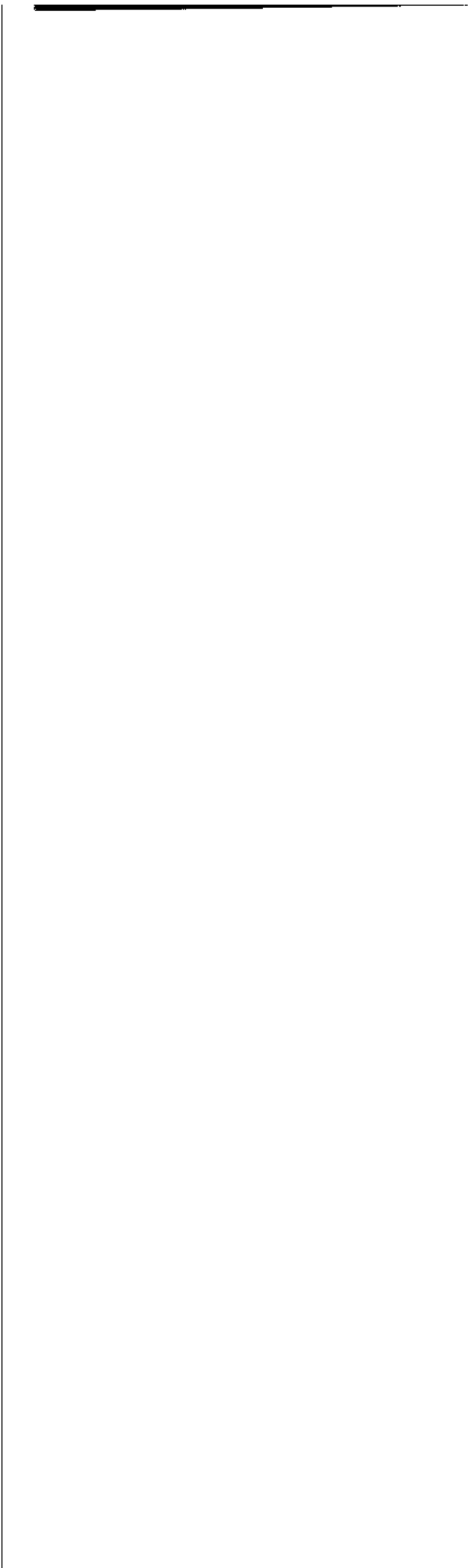
这并不足以证明用环丙烷比其他麻醉药危险。举出有哪些潜在变量可能和手术时麻醉药的选择发生交叉，因此可能可以解释不同的死亡率。

5.24 随机化的执行。为了示范随机化怎样可以减少交叉，考虑以



下情况。一位营养师通过实验试图比较，刚断奶的公鼠喂食饲料 A 或饲料 B，体重增加的情况。她会用每种饲料喂 10 只老鼠来比较。她有 10 只老鼠属于遗传品系 1，另 10 只属于品系 2。品系 1 体格较强壮，所以如果品系 1 的 10 只老鼠都喂饲料 A，则品系和饲料的效应会混杂在一起，实验结果会有偏差，偏向对饲料 A 有利。

- (a) 把老鼠编代码 00、01……19。用表 A 来指派 10 只老鼠给饲料 A。一共做 4 次，每次用表 A 的不同部分，并写下被分派到饲料 A 的 4 个组的成员代码。
- (b) 实验者并不知道，代码 00、02、04、06、08、10、12、14、16 和 18 的 10 只老鼠，属于遗传品系 1。刚才产生的 4 个饲料 A 组中，每组各有几只品系 1 老鼠？品系 1 被指派到饲料 A 的老鼠平均有几只？



第 6 章

真实世界中的实验

多变的小鼠

随机化比较实验可能是统计学里最重要的观念。要证明改变一个变量真正会导致另一个变量改变的话，实验是提供证据的金字招牌。然而在真实世界中，实验并不是无往不利的。即使实验过程顺利，我们也不一定可以对实验结果坚持到底，看看多变小鼠这个例子就可以知道。

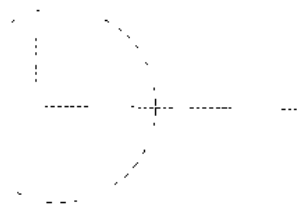
我们的行为(一部分)由基因决定，是目前的重要科学议题之一。我们不能拿人来做实验，所以就用小老鼠代替。研究者将一组小鼠“敲掉”一个基因，然后比较这组小鼠和控制组里正常小鼠的行为。在执行处理之前，所有小鼠的基因组成是一样的，而且是用随机方式



指派到处理组或控制组的。这是一项毫无漏洞的随机化比较实验：如果两组行为有别，则必定是受了被敲掉的基因的影响。可是呢，却是像《科学》期刊一位作者说的：“一组研究者刚刚才做出结论，把某种行为归因于某个基因，马上就接着有另一项研究，证明前一项研究的结果不正确，甚至认为那个基因恰恰有和前项结果中所说的相反的效应。”

这不仅让科学家很沮丧，也让记者泄气，因为他们想要报道，“科学证明”我们的种种行为，全是由基因决定的。到底出了什么差错呢？

有些沮丧的科学家尝试找答案。他们用同一个品系的小鼠，在三个不同的实验室，做了相同的实验。一位研究者形容，为了要让分别在俄勒冈、加拿大艾伯塔省及纽约的实验室，在各种条件方面都一样，他们都差点要“发狂”了。可是得到的结果常常还是很不一样。似乎只要实验室的环境有小小的不同，就会对小鼠的行为造成很大的影响。下次谈到基因控制我们行为的时候，别忘了这一段内容。



一视同仁

概率样本是重要概念，但是实际抽样时会发生困难，而这并不光是用随机样本就可以解决的。随机化比较实验也是重要概念，但是也没解决实验时发生的所有问题。抽样的人必须完全清楚他要什么样的信息，并且把问题写得非常明确，使他能从样本当中汲取所要的信息。而做实验的人则必须确实知道，他要的是哪些处理和哪些反应的资讯，并且他必须能够构建出实施处理方法和度量反应所必需的装置。这就是当心理学家、医学研究者或工程师说到“设计实验”时的实际意思。我们关心的是实验设计的统计面，这些统计观念对于心理学、医学、工程及其他领域的实验都同样适用。即便我们讨论的层面并不深入，还是应该要了解存在哪些实际问题，会使得实验无法产生有效的数据。



随机化比较实验背后的逻辑是，对所有的受试对象在各方面都一视同仁，惟一的不同就是那些在实验中设计出的，用来做比较的处理方法。在任何其他方面有不同的对待就会产生偏差，但是要对所有受试对象在所有方面一视同仁，是很困难的任务。

例1 小鼠、大鼠及兔子

许多实验中的受试物是特别繁殖的小鼠、大鼠和兔子，这些动物经过特别繁殖，因而拥有相同的遗传特性。就像多变小鼠的例子中说到的，动物和人一样，对于怎样被对待可能相当敏感。以下是两个关于不同对待方式会如何制造偏差的有趣例子。

某种新的早餐谷片有没有营养呢？为了找出答案，就喂一些大鼠吃新产品，另一些大鼠吃标准食物，然后比较两组的体重增加情形。大鼠被随机指派吃哪种食物，而且住在架叠起来的笼子里。结果发觉，住在上层笼子的大鼠长得比住在下层的快些。如果实验者把吃新产品的大鼠放在上层，而把吃标准食物的大鼠放在下层，则这个实验就会有偏差，有利于新产品。解决方法：将大鼠随机指派到笼子里。

另外有个研究是想看看人类的情感对兔子的胆固醇浓度有没有影响。所有受试兔子都吃一样的食物，但某些兔子(随机选取)会定时被放出笼子，让一些友善的人搔它们毛茸茸的头。结果，受到关爱的兔子胆固醇浓度较低。所以，在以兔子的胆固醇浓度为反应变量的实验里，如果只对某些兔子表达关爱，而没有遍及其他兔子，就有可能对结果造成偏差。

双盲实验

安慰剂有作用。仅仅因这项事实，医学研究就必须特别费点事去证明，一项新疗法并不仅是安慰剂而已。对所有受试对象一视同仁的部分原因，是为了要确定安慰剂效应已作用在每一个受试对象上面。



例2 强有效的安慰剂

想帮开始秃头的男性保住他们的头发吗?给他们安慰剂就成了!有一项研究发现,一些秃头男性在服用了安慰剂之后,有42%的人脑袋上的头发保住甚至增加了。另外一项研究对13个对野葛敏感的人说,涂在他们一只手臂上的东西是野葛,而其实那是安慰剂,但是13个人全部起了疹子。事实上,涂在另一只手臂上的才真的是野葛,但是受试对象被告知那是无毒的——结果13个人当中只有两个人起疹子。

对于比较不明确的疾病,而且是心理层面的,比如说忧郁症,有些专家认为,最常用的药物中约有四分之三,其效用不过是安慰剂效应罢了。但这个说法有些人并不同意。医学实验中安慰剂效应的强度很难确定,因为就像多变小鼠的行为那样,实验环境对实验结果有很大的影响。而医师是否很积极,也有非常大的影响。不过当你考虑要计划一项医学实验时,先想到“安慰剂有作用”,会是一个好的开始。

安慰剂效应的强度,是随机化比较实验的有力论据。在秃头实验中,安慰剂组有42%的人保住或增加了头发,但是在使用一种新的防秃药的那一组,有86%的人保住或增加了头发。防秃药打败安慰剂,所以代表防秃药的效用,不只是安慰剂效应而已。当然安慰剂效应仍然是这种药以及其他疗法有效的部分原因。因为安慰剂效应这么强,所以如果告诉医学实验的受试对象他接受的是新药抑或是安慰剂,可就有点笨了。如果他们知道自己得到的“只是安慰剂”,可能会减低安慰剂效应,使得实验结果有偏差,偏向对其他疗法有利。如果告诉医师或其他医护人员,每个受试对象所得到的处理是什么,那也一样不妥。如果他们知道某个受试对象得到的“只是安慰剂”,他们的期望就会比“知道受试对象得到的是有希望的实验药时”来得低。医师的期望会改变他们和病人的互动,甚至影响他们对病人病况的诊断。因此只要有可能,任何以人当受试对象的实验,都应该做到双盲。



统计学上的争议

到底是不是安慰剂?

自然疗法是大生意：鹿茸可以增强运动能力；缬草提取物可减缓压力、头痛及经痛；育亨宾对性生活有帮助。商店的货架以及网站上充斥了各式各样号称对你健康有帮助的奇特物品。

没有在随机实验中和安慰剂做过比较的疗法，本身就可能只是安慰剂而已。美国法律要求，新的处方药及新的医疗装置，必须经过随机化试验来证明其安全性及有效性。因此你可以有信心，你的医师开给你的药不会只是安慰剂而已。

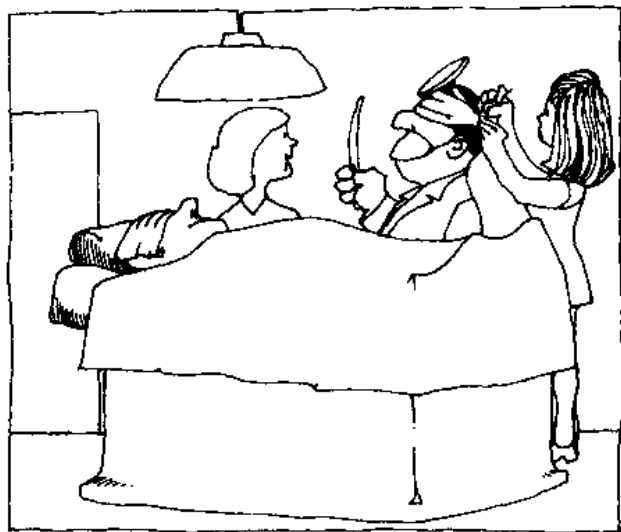
至于那些“自然疗法”又如何呢？美国法律允许药草、维生素及营养品的制造商，不用证据就可以声称这些产品：安全而且会增强“自然条件”。他们不能宣称可治疗“疾病”。当然“自然条件”和“疾病”之间的界限很模糊。

我不用任何证据就可以说，墨尔博士的“印第安纳陈年菁华液”可以促进心脏健康。没有经过临床试验和美国食品和药物管理局 (Food and Drug Administration) 的认可，我不能说它可以减低心脏病的风险。但是，当人家看到我的广告时，无疑有很多人会把“促进心脏健康”和“减低心脏病的风险”当做同一回事。我也不必担心我的药丸

里含的“印第安纳陈年菁华液”的剂量是多少，或者是怎样的剂量会有害。

食品和药物管理局是不是应该要求自然疗法也要遵守和处方药一样的标准呢？实际上很难这样要求，因为自然物质不能申请专利。制药公司可以花几百万在临床试验上面，而一旦证明药物有效，他们就可以申请专利。没有人可以替草药申请专利，因此也就没有人愿意花钱去做临床试验。所以别指望目前的规定会有什么大幅度的改变。

目前的情况是，你很容易就会听到诸如银杏有益于(如某一个网站所说的)治疗“听觉和视觉问题、阳痿、水肿、静脉曲张、中风及腿部溃疡”这类的声明。用常识判断就知道，不管任何东西，如果号称能对各种不相关的问题都有效的话，就应该对此表示怀疑。统计知识告诉你，应该要对背后没有比较实验在支持的声明存疑。许多未经过测试的疗法，无疑只是安慰剂而已。然而它们可能对许多人真的有疗效，因为安慰剂效应是很强的，但是要记住，人家配制出来的这些东西，安全性可是没经过检验的。



“伯恩斯医师，您确定统计学家说的双盲实验是这个意思吗？”

• 双盲实验

在**双盲实验**(double-blind experiment)当中，不论是受试对象，还是会和受试对象有互动的人，都不知道哪位受试对象接受了哪种处理。

直到研究结束，结果出来为止，只有该研究的统计学家知道全部情况。我们以一项对鼻喷剂形式的流行性感疫苗做的研究当例子，可以发现医学期刊中的报告通常是如此开始的：“这项研究是随机化、双盲而且有安慰剂控制组的试验。参加者是在1997年9月中旬到11月中旬之间，在遍布美国大陆的13个地点登记加入的。”医师都应该知道这些话是什么意思，现在你也知道了。

拒绝参加、不合作者及退出者

抽样调查有“无回应”的问题，原因是联络不到样本中的某些人或有些人不愿意回答。用人做受试对象的实验也有类似的困扰。

参加实验却不遵循实验处理的受试对象叫做**不合作者**(nonadherer)，不合作者也可能造成偏差。举例来说，参加新药试验的艾滋病人有时会自己加上其他的治疗。还不只这样，有些艾滋病人把他们的药拿去化验，如果发觉自己不是分配到新药组，就会退出或自己加其

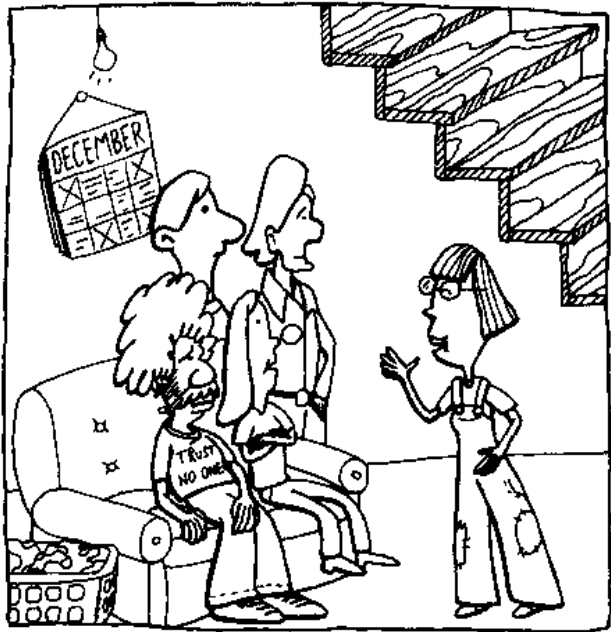


例3 医学实验中的弱势群体

严重疾病(比如癌症)疗法的医学实验中,受试对象的拒绝参加已成为严重问题。跟抽样时的情况一样,如果拒绝参加的人和愿意合作的人之间有系统性的差异,就可能造成偏差了。

少数族裔、女性、穷人以及老人,长期以来在临床试验中的代表性都不足。很多时候都是没人找他们参加。现在法律已有规定必须包括女性和少数族裔,从数据中可以看出,现在大部分的临床试验,女性和少数族裔已获有公平的代表性。但是拒绝参加仍然是问题,其中少数族裔,尤其是黑人,不愿加入的机会最大。政府相关主管单位卫生部少数族裔办事处(Office of Minority Health)说:“虽然最近一些研究显示,非洲裔美国人对于癌症研究的态度渐趋肯定,还是有些研究结果确认,他们对临床试验仍然有疑虑。使他们不愿参加的主要阻碍,是对于医疗机构的不信任。”对于缺乏信任的补救措施,包括提供实验的完整而清楚的资讯,将正在实验的处理投保,有黑人研究者的参与,以及在黑人社区中和医师及医疗组织合作等。

实验捣蛋联盟会议



“长江一号,你负责摇实验兔子的头;长江二号,你要设法混入临床试验小组,然后半颗药也不要吃;长江三号,你想办法去参加一个实验,然后中途逃跑。”

他的药。这样会造成对新药不利的偏差。

持续时间较长的实验也常碰上退出者(dropout),也就是开始时参加实验却不完成实验的受试对象。如果退出的原因与实验的处理无关,则没什么妨碍,只是受试人数减少罢了。如果受试者退出是因为对某个处理的反应,就可能造成偏差。

我们的结论可以推广吗?

设计完善的实验可以告诉我们,解释变量的改变造成了反应变量的改变。说得更明确一点的话,它告诉我



们的结论，是在这个特定实验的特定环境之下，发生在特定受试对象身上的事。但我们希望的不仅如此。我们希望能够声明，我们教数学

例4 医学实验中的退出者

“赛尼可”是一种新药，它会阻碍我们从食物中吸收脂肪，因此可能可以帮助我们减肥。像一般程序一样，这种药经过一项双盲随机化试验，和安慰剂做了比较。以下是比较的过程。

先从1187位肥胖的人开始，先给4个礼拜的安慰剂，然后把不愿按时服药的人剔除。这样做是先把不合作者做了初步过滤，如此剩下892位受试对象。把这些人随机指派到赛尼可组或安慰剂组，并为他们设计了减肥餐。此减肥计划开始一年后，还有576位受试对象仍然继续参与。平均来说，赛尼可组比安慰剂组多轻了3.15千克(约7磅)。计划又继续了一年，这一年的重点是保住前一年已减掉的体重不要回升。第二年结束时，还剩下403位受试对象。这只是刚开始随机化时的892个人的45%而已。赛尼可又打败安慰剂，回升体重平均来说少了2.25千克(约5磅)。

有那么多受试对象退出时，结果还可靠吗？两组的整体退出率很接近：赛尼可组是54%，安慰剂组是57%。退出的原因和处理有关系吗？减肥实验中的安慰剂组受试对象，通常因为体重没有减少而退出。这点会使研究偏向对赛尼可不利，因为研究结束时仍在安慰剂组的人，可能是那些只要肯持续吃减肥餐就能减肥的人。研究者详细检视了退出者的所有资料。两组的退出者，减轻的体重都比留下的人要少，但是经过仔细的统计分析，认为偏差很小。也许这是事实，但是这样的结果就比较不明确，不像我们原先认为实验应该达到的结论。

的新方法，能使一般中学生都学得更好；或者我们的新药对于众多病人来说，都比安慰剂有效。究竟我们能不能把我们的结论，从一小组受试对象，推广到广大群众呢？

首先要确定的是，我们的结论有统计显著性，也就是说证据强到很少会光靠机遇而发生。这件事很重要，但它是技术上的细节，参与研究的统计学家会帮我们确认这一点。比较严重的威胁是，实验中的处理、受试对象或者实验环境也许不切实际。我们来看看一些例子。



例 5 挫折的研究

一位心理学家想研究，失败和挫折对于一个工作小组成员间的关系有何影响。她将学生组成一队，带他们到心理实验室，然后叫他们玩一种需要团队合作的游戏。游戏被做了手脚，使得他们总是输。心理学家透过单向窗，观察这些学生玩一晚上的游戏，并且记下他们的行为变化。

在实验室里玩马上就会结束且赌注很小的游戏，比起工作好几个月开发新产品结果总是有问题，最后被公司放弃，可差了十万八千里。学生在实验室里的行为能提供我们多少信息，让我们了解“一个产品失败的工作小组的行为”呢？

如果心理学家的目标是要对“职场中的挫折对于团队工作的影响”做出结论的话，那么例 5 当中的受试对象(知道自己是一项实验中受试对象的学生)，处理(做了手脚的游戏)，以及环境(心理实验室)都很不切实际。虽然心理学家尽量设计比较切合实际的实验来研究人的行为，但是实验和实际情况的差距仍然使得这个领域的实验用处有限。

例 6 第三煞车灯

1986 年开始，在美国出售的汽车除了汽车尾部原有的两个煞车灯外，还必须在中央高处加装第三煞车灯。通过出租车及商务用车的随机化比较实验证实，这个安全要求是有道理的。实验显示，第三煞车灯将车尾碰撞减少了五成之多。

施行了近 10 年之后，美国的保险协会(Insurance Institute)发觉，汽车尾部碰撞只减少 5%，因此这项规定虽然有帮助，但和实验的预测差了很多。这是怎么回事？当年执行实验时，大部分车都还没有第三煞车灯，所以第三煞车灯很容易吸引后方驾驶的视线。现在几乎所有车都有第三煞车灯，第三煞车灯也就不再引人注意了。实验结论推广得不如安全专家所希望的那样好，是因为环境已经不一样了。



例7 受试对象是不是受到太好的待遇？

医学实验应该切合实际了吧，毕竟受试对象都是真正的病人，在真正的医院里，因为真正的疾病而接受治疗。

不过即使在这种状况下，仍然有一些问题。参与医学试验的病人得到的医疗照顾，比大多数其他病人来得好，即使是被分到安慰剂组的亦然。他们的医师是研究此种病症的专家。他们比其他病人得到更多照顾。因为一直有人提醒，所以他们也更会按时服药。除了实验疗法和控制疗法的不同之外，对所有病人提供“相同的照顾”，其实等于说“对所有病人提供最好的医疗照顾”。结果就是：当新疗法提供给一般病人使用的时候，效果也许不如在临床试验中的受试对象那样好。在临床试验中胜过安慰剂的疗法，正式使用时多半也是会赢过安慰剂，只是试验结果中的“治愈率”或其他度量成功程度的数字，可能会偏于乐观。

卡罗来纳启蒙计划(第5章)也面对了同样的“结果太好了，不太可能是真的”的问题。这项既费时又费钱的实验的确显示出，密集的日间照护，对儿童的日后生活有很大的好处。研究中所做的日间照护的确够密集：有许多优秀的工作人员，许多家长的参与，以及儿童从很小就开始参加精密策划的活动，全部的花费差不多是每个儿童每年11 000美元。我们的社会不大可能会决定给所有低收入儿童这样的照顾。有个大问题还没有答案：日间照护要好到什么程度，才能真的帮助儿童在未来成功？

在实验不完全切合实际时，实验数据的统计分析没法告诉我们，结论可以推广到什么程度。实验者如果把结果从实验室中的学生身上，推广到真实世界中的上班族身上，必须能够根据他们“对人如何运作的了解”来说服大家，而不是只根据数据。要从实验室的老鼠推广到真实世界的人身上，就更难上加难了。因此，即使实验设计的逻辑非常有说服力，单单一个实验也极少能使人完全信服。新发现常常必须在不同背景下经过多次实验的探讨，才能找到真正的适用范围。

实验是否已实际到可以产生有用的信息？能否让人信服？这不是根据统计理论决定，而是由实验者对实验主题领域的知识的掌握情况决定。要避免产生隐性偏差所需注意的一些细节，也依赖于对主题领域的



后设分析

对一项重要议题只做了一次研究，很难就此下结论。通常会有好几项研究，它们的背景不同，设计不同，品质也不一样。我们能不能把不同研究的结果整合在一起，当做一个整体的结论呢？这就是“后设分析”（meta-analysis）背后的概念。当然啦，各个研究之间的差异，使得我们无法直接把结果凑在一起。统计学家会用较深奥微妙的方法来整合结论。后设分析曾被用在二手烟的影响，以及补习是否可以增进SAT考试的分数这些议题上。

知识的了解。好的实验须结合统计原则及对研究专业领域的了解。

真实世界中的实验设计

我们已见过的实验设计都遵循同样的模式：先把受试对象随机分组，组数和处理数相同，然后对每一组施行一种处理。这些叫做完全随机化设计（completely randomized design）。

完全随机化设计

在完全随机化（completely randomized）的实验设计中，所有的实验个体都随机配置给所有的处理。

还不只如此，到现在为止我们的例子当中都仅有一个解释变量（真药对应于安慰剂，教室学习对应于网络学习）。而完全随机化设计可能有任何数目的解释变量。以下举的就是有两个解释变量的例子。

例 8 电视广告的效果

重复看到同一个广告的效果如何？答案可能和广告长度以及播出频率都有关系。有一项实验用大学生当受试对象，探讨了这个问题。所有受试对象都观赏了一个 40 分钟长的电视节目，其中包含某种数码相机的广告。有些人看了 30 秒的广告，其他人看的是 90 秒的版本。同一个广告在该电视节目中出现 1 次、3 次或 5 次。节目结束后，所有受试对象都要回答问题，问题关于他们对广告内容的印象，他们对该相机的好恶，以及是否有购买意愿。这些都是反应变量。

这项实验有两个解释变量：广告长度（分成 2 种等级）以及重复次数（分成 3 种等级）。两个变量的各一种等级搭配起来，共有 6 种组合，构成了 6 种处理。图 6.1 显示出处理的配置（layout）。



		变量 B(次数)		
		1 次	3 次	5 次
变量 A(长度)	30 秒	处理 1	处理 2	处理 3
	90 秒	处理 4	处理 5	处理 6

图 6-1 例 8 中实验的处理。两个解释变量共搭配出 6 种处理

实验者常常会想同时研究好几个变量的联合效应(combined effect)。几个因素的相互作用(interaction)所产生的效应,无法从每个因素的单独效应预测出来。或许长一点的广告会使观众对产品的兴趣增加,多播几次广告也许会增加兴趣,但是若我们既将广告加长,又多播几次,观众可能就会厌烦了,对产品的兴趣也随之减低。例 8 里的实验会帮我们找出答案。

配对及区集设计

完全随机化设计是统计实验中最简单的,而且这类设计清楚描述了控制及随机化这两项原则,不过完全随机化设计常常不如一些更复杂的统计设计。确切一点说,用各种方式将受试对象做一些适配,得到的结果比起只做随机化还更精确。

结合适配和随机化的常用设计就是**配对设计**(matched pairs design)。配对设计只比较两项处理。先选取成对的受试对象,同一对中的两个要尽量接近。然后利用掷铜板方式,或者从表 A 中读出的随机数字为奇数或偶数来决定,把二个处理分别指派给每一对当中的两个受试对象。有时候配对设计中的“一对”,实际上只包含一个受试对象,只是分时间先后分别接受两个处理。此时每个受试对象就是他/她自己的控制组。接受处理的顺序可能影响受试对象的反应,所以会再用铜板来随机化每个受试对象接受处理的顺序。

配对设计当中用到了比较处理以及随机化两项原则。不过此处的随机化并不是完全随机化,因为我们并没有同时把所有受试对象



都随机指派给这两个处理。我们做的只是在每一个配对中随机化而已。这样的做法可经由配对，减少受试对象间的变异所产生的影响。而配对设计是区集设计(block design)的一个特例。

例9 可口可乐对百事可乐

百事可乐想要证明，可口可乐的爱好者在两种可乐都不予标示下进行试喝时，事实上会偏爱百事可乐。受试对象全是自称嗜喝可口可乐的人，他们从没有标示品牌的玻璃杯当中喝了两种可乐之后，要说出比较喜欢哪一种。这就是配对设计：由每一个受试对象比较两种可乐。因为反应可能会和先喝哪种可乐有关，所以每个受试对象喝两种可乐的顺序都应该随机选择。

当超过一半的可口可乐爱好者选择了百事可乐时，可口可乐声明，该项实验有偏差。因为放百事可乐的玻璃杯有M的记号，可口可乐的有Q的记号，可口可乐就说，你看，这只代表大家喜欢M这个字母超过Q。配对设计本身没有问题，不过比较严谨的实验，会避免在可口可乐和百事可乐本身的差异之外，还有其他的差别。

• 区集设计

一个区集(block)就是一组实验个体，这些个体在实验之前，就被认为在会影响反应的某些方面很类似。区集设计当中，将个体随机指派到各处理的这个步骤，是在每个区集里面个别执行的。

区集设计结合了用适配来制造相近的处理组的概念，以及用随机方式产生各处理组的原则。区集是另一种形式的控制：通过把外在变量引进实验里来造成区集，可以控制这些外在变量的影响。以下是区集设计的一些典型的例子。

区集是在实验开始之前就先分好组的受试对象。我们把“处理”这个名词特别保留，来当作我们加在受试对象上的条件。虽然在例10当中，我们可以比较由2个区集(男性、女性)及3种广告形成的6组人的反应，我们却不把那6组叫做6个处理。区集设计类



例 10 男性、女性和广告

女性和男性对广告的反应不一样。有个实验要比较同一产品的3个电视广告的效率,除了对这些广告的整体反应外,也想要知道男性和女性的反应如何。

完全随机化设计会把所有受试对象,包括男性和女性,全部放在一起考虑。

“随机化”的部分是把受试对象分派到3个处理组,完全不理睬他们的性别。但这样做等于将男性和女性的差别置之不理。比较好的设计是将男性和女性分开考虑:将女性随机指派到3个组,每组看一个广告,再将男性也随机指派到3个组。图6.2描绘了这个经过改良的设计。

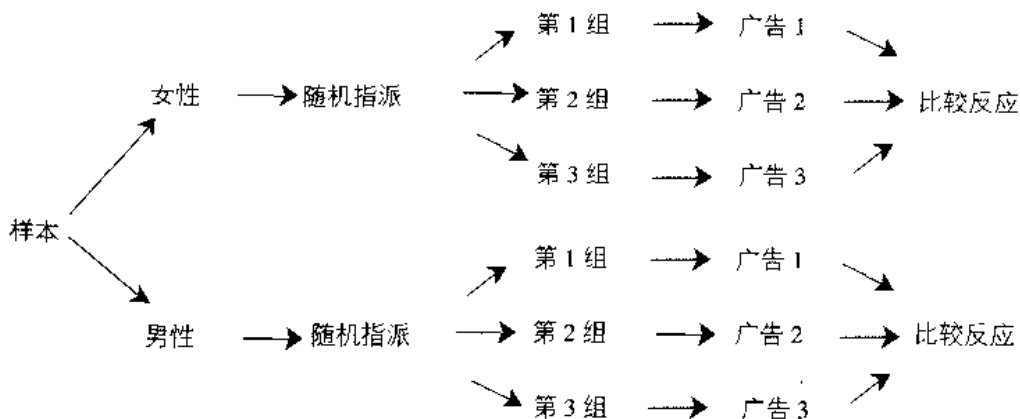


图 6.2 用来比较3个电视广告效果的区集设计。男性受试对象和女性受试对象构成两个区集

例 11 比较福利政策

某一项社会政策实验,想要评估几个新提出的福利制度对家庭收入的影响,并和现存的福利制度做比较。因为一个家庭的未来收入和它目前的收入密切相关,所以愿意参加这项实验的家庭,就依照收入的多寡,分成几个区集,收入接近的分在同一个区集。然后同一区集中的家庭,再随机分派到不同的福利制度。



似抽样中的分层样本。区集和层都是把近似的个体聚集起来。我们用不一样的名称，只是因为这个概念是在抽样和实验两方面分别发展出来的。区集设计的优点和分层样本一样，有了区集，可以让我们分别对每个区集做结论；比如说，在例 10 的广告研究中可以分别对男性及女性做结论。分区集也可以使得整体结论更精确，因为当我们研究 3 个广告的整体效果时，可以把男性和女性之间系统性的差异去除掉。区集的观念，是统计实验设计中另一个重要原则。明智的实验者会根据实验受试对象之间最重要且无法避免的差异来源，来组织区集。然后随机化会把剩下变异的效应给加加减减平均掉，而使得处理之间能有个无偏的比较。

就像设计样本时一样，要设计复杂的实验也是专家的事。在我们已经对实验的有关知识略知皮毛之后，我们通常还是会假装，大部分的实验都是完全随机化实验。

网络寻奇

美国卫生部少数族裔办事处的新闻通讯《把差距缩小》(closing the Gap) 有一期特别谈临床试验。这一期里有好几篇有关少数族裔参与医学实验的文章，你可以在 www.omhrc.gov/ctg/ctg~ct.htm 网站找到。

你也可以在《郎中现形》(Quackwatch) 网站找到一些没有经过合适的临床试验支持，和健康有关的(非常)可疑的声明。网址是 www.quackwatch.com。



本章重点摘要

和样本一样，实验也需要有好的统计设计，加上一些对付实际问题的方法。因为**安慰剂效应**很强，所以**临床试验**及其他以人当作受试对象的实验，只要可能都应该执行**双盲**。双盲方法可以有助于达成**比较实验**的基本要求：除了实验要比较的处理之外，其他方面**对所有受试对象一视同仁**。

如同**抽样时有无回应**的问题一样，做实验也会碰到**不合作的受试对象**。有些人拒绝参加；有些人在实验未完成前就退出；还有些人不遵照指示，比如有些人在药品实验中不按规定服药。实验最常见的弱点是，我们不能将得到的结果**普遍推广**。有些实验用了**不切实际的**处理，有些用的受试对象是从**特定群体**中选出的，比如只从大学生中选出，而且所有实验都在**特定场所和特定时间**执行。我们会希望，类似的实验在**不同的地点和不同的时间**执行，以确认我们的**重要发现**。

许多实验用的设计，比基本的**完全随机化设计**还要复杂。完全随机化设计是把所有受试对象，随机分配给所有的处理。**配对设计**比较两个处理，方法是把两个处理分别给一对类似的受试对象，或者两个处理先后给同一受试对象，但顺序则随机决定。**区集设计**先把类似的受试对象放在同一个区集，然后分别在每一个区集中随机指派处理给各受试对象。令人信服的实验，关键就在于**随机化、控制以及足够数目的受试对象**这些重要概念上。



第6章 习题

6.1 医药新闻。当研究发现羟基脲可以缓解地中海贫血的症状时，美国国家健康研究所发生了一份医学报告。报告中说：“这些发现从一项研究中数据分析的结果而来，这项研究是针对地中海贫血做的羟基脲多通道研究(MSH, Multicenter Study of Hydroxyurea)，是一个双盲且有安慰剂控制组的试验，受试对象中一半接受羟基脲，一半接受安慰剂胶囊。”向不懂统计的人解释，这里的“有安慰剂控制组”及“双盲”是什么意思。

6.2 冥想是否能减低焦虑？一项声称已证实冥想可减低焦虑的实验，是以下述方式进行的。实验者先和受试对象面谈，并把他们的焦虑程度分级，然后把受试对象随机指派分成两组。实验者教其中一组如何冥想，之后他们每日进行冥想并持续一个月。对于另一组，实验者只叫他们放轻松而已。一个月后，实验者再和所有受试对象面谈，并替他们的焦虑度分级。冥想组的焦虑程度比之前要低。心理学家说结果有点可疑，因为在评估焦虑程度时并不是“盲目”的。说明这是什么意思，以及为何不盲目可能会使结果有偏差。

6.3 急诊室医疗。某医学期刊中有篇文章报道了一项实验，实验目的是要了解：如果中风病人除了进行急诊室标准程序之外，再多注射一种含氧液体，是否可以减轻病人的出血状况？文章中描述该实验为“随机化、受控(controlled)、单盲效能试验(single-blinded efficacy trial)，从1997年2月到1998年1月之间，在美国18个创伤中心执行。”你认为这里的“单盲”是什么意思？为什么不可能执行双盲实验？

6.4 日间行车灯。加拿大规定汽车必须有“日间行车灯”，也就是车子一发动就会自动亮起来的低亮度头灯。有些汽车制造商已开始在美国出售的车上也装置行车灯。有没有可能，因为行车灯提高汽车对于其他驾驶员的“能见度”，而减少事故的发生呢？

(a) 简单描述一个可以帮助解答这个问题的实验设计。你会检视哪些反应变量？

(b) 例6讨论到第三煞车灯，从那个例子中你学到的注意事项，有哪



些是可以应用在行车灯效应实验上的?

6.5 食物及癌症 致癌物不应该出现在我们的食物里。我们不愿在人身实验,看什么东西会致癌,所以我们用大鼠来实验。大鼠是特别饲养的,使其的肿瘤比人类多,这些大鼠在长约两年的一生中,大部分时间都被喂食高剂量的实验化学物质。大略讨论一下,如果用这些实验来决定哪些东西加在人类食物里是安全的,会有些什么问题?

6.6 胆固醇问题 临床试验曾证明,不论用药物还是食物来降低血液中的胆固醇含量,都会减低心脏病的风险。最早做的一些研究,追踪了受试对象5—7年,为了要在这段不够长的时间当中看到结果,受试对象是从最高危险群选出的,也就是有高胆固醇或者已有心脏病的中年男性。实验大致都得到降低胆固醇可以降低心脏病风险的结论。有些医师质疑,这些实验结果对他们的病人是否都能适用。为什么?

6.7 检验自然疗法 虽然法律没有强制规定,我还是决定要让墨尔本博士的“印第安纳陈年菁华液”接受临床试验检验。我想要证明该菁华液可以减轻关节炎造成的疼痛。有60位需要止痛剂的关节炎患者可以参与实验,我曾给每位患者1粒药丸,1个小时后再问:“你觉得你的疼痛缓解了几成?”

- (a) 为什么我不应该把菁华液给全部60位病人,再记录他们的反应?
- (b) 描述实验设计,以比较菁华液、阿司匹林及安慰剂的效果。
- (c) 应不应该告诉病人他们得到的是什么治疗?如果病人知道后,可能会对他们的反应有怎样的影响?
- (d) 如果病人未被告知接受的是什么处理,这项实验叫做单盲。这项实验是不是应该设计成双盲呢?请解释。

6.8 检验自然疗法 美国国家健康研究所终于提供经费,针对某些自然疗法做了适当的临床试验。其中有一项在杜克大学(Duke University)进行的研究中,330位有轻微抑郁症的病人参加了一项试验,比较贯叶连翘(St. John's wort)、安慰剂以及左洛复(Zoloft)的效用,左洛复是抑郁症的常用处方药。贝克抑郁问卷(Beck Depression Inventory)是常用来评估抑郁症严重程度工具,评估等级从0到3。

- (a) 你曾想用怎样的反应变量,来度量经过治疗后抑郁症的变化?
- (b) 为这项研究设计一个完全随机化临床试验,描述概要即可。



(c) 在这个试验中还要注意什么事项?

6.9 安慰剂效应。对一些医师做的一项调查发现,有的医师对于抱怨疼痛却又找不出原因的病人给予安慰剂。若病人的疼痛减轻,这些医师即推断这种结果毫无实质的根据。进行这项调查的医学院研究者声称,这些医师不了解安慰剂效应。为什么?

6.10 膝伤低痛治疗法。受伤的膝盖现在可以做关节镜手术,并不需要把膝盖切开。如果给病人非类固醇消炎止痛药(NSAID, nonsteroidal anti-inflammatory),能够减轻病人的不适吗?83位病人被分成3组。A组在手术前后都服用了NSAID。B组在手术前得到的是安慰剂,手术后才是NSAID。C组在手术前后都得到安慰剂。病人在手术一天后回答问题时记录下“疼痛分数”(pain score)。

(a) 描述这项实验设计。你不需要实际执行设计中的随机化部分。

(b) 你读到在这项研究中“病人、医师及物理治疗师都是盲目的”。这是什么意思?

(c) 你还读到“A组的疼痛分数有统计显著性地低于C组,但并非有统计显著性低于B组。”这是什么意思?这项发现会让你对使用NSAID做出什么结论?

6.11 比较玉米品种。有不同氨基酸含量的新品种玉米,也许营养价值高于传统玉米,后者的赖氨酸(氨基酸的一种)含量较低。有一项实验将两种新品种,叫做不透明2号(opaque-2)及粉质2号(floury-2),和正常玉米比较。研究者把玉米和大豆混合做成饲料,每一种玉米都有3种不同的蛋白质含量,分别为12%、16%以及20%。他们把每一种饲料喂给10只1天大的小公鸡吃,并在21天之后记录小公鸡增加的体重。公鸡增加的体重,是所吃饲料营养价值的一种指标。

(a) 这个实验中的个体及反应变量分别是什么?

(b) 一共有几个解释变量?几个处理?用像图6.1的图来描绘这项实验。一共需要多少个实验个体?

(c) 用一个图来描绘这项实验的完全随机化设计(不必真的执行随机化)。

* 译注:材料已配好,只要烤熟就可以吃。

6.12 烤蛋糕。某家食品公司想要新上市一种蛋糕半成品*。重要的事情是,当烤的时间或者温度有少许不同时,蛋糕的味道不应该不一样。



在一项实验当中,他们把这种半成品分别用 300°F、320°F 及 340°F 三种温度烤 1 小时或 1 小时 15 分钟两种时间,用每一种温度和时间的组合烤了 10 个蛋糕。再找一些试吃者给每个蛋糕的口感和味道打分数。

(a) 这个实验的解释变量和反应变量是哪些?

(b) 画一个像图 6.1 的图来描述处理。一共有几个处理?总共需要多少个蛋糕?

(c) 假如先针对一个处理同时烤 10 个蛋糕,再对第二个处理同时烤 10 个蛋糕,以此类推,如此做不算是好主意,为什么?实验者的实际做法,是根据设计中的随机化部分,随机决定烤蛋糕的顺序。

6.13 比较两手的力量大小。习惯用右手的人,是否右手比左手有力?手力大致可以这样量:把家庭用体重计放在架子上,一端突出在架子外,然后拇指在下、另四指在上挤压体重计,体重计的读数显示受力的强度。用 10 位惯用右手的人当做受试对象,描述可以比较左手和右手力量的配对实验的设计(不必实际执行随机化)。

6.14 走势图对投资者有帮助吗?有些投资者相信,过去股票价格的走势图,对预测未来价格有帮助,对此大部分经济学家都不同意。在一项检验走势图效用的实验当中,商学院的学生在电脑屏幕上(假设性的)买卖外币。一共有 20 位学生参与,为方便姑且称之为 A、B、C……T。他们的目标是尽量多赚钱,表现最好的几位有奖品。在学生外汇操作员的电脑里,有过去外币兑美元的价位资料,其中有些人的软件还可把趋势强调出来。针对这项实验描述两种设计,即完全随机化设计以及配对设计,在后者中以每位学生当做自己的控制组。在两种设计中都要实际做出随机化的部分。

6.15 温度和工作表现。研究工作表现的专家想要知道,对于需要用手处理的工作,室内温度会不会影响工作表现,她选择了 70°F 和 90°F 两种温度当作处理。实验装置是一种“栓和洞”的装置,必须同时用两只手完成将木栓塞进洞中的动作,反应变量就是 30 分钟内正确塞进洞中的木栓个数。每个受试对象先在该装置上训练,然后就被要求在 30 分钟内连续不停,尽量多塞几根木栓。

(a) 描述一个完全随机化设计,来比较 70°F 和 90°F 时手的灵巧度,共有 20 位受试对象参与实验。

(b) 因为人和人之间,手的灵巧度可能有很大的差别,个别结果之间



的大变异，有可能把温度的系统效应给模糊掉了，除非每组都有很多受试对象。详细描述一个配对实验的设计，把每个受试对象当做自己的控制。

6.16 比较癌症疗法。某种癌症的进展速度男女有别。因此一项比较比癌症三种疗法的临床实验，就把性别当做区集变量 (blocking variable)。

- (a) 有 500 位男病人和 300 位女病人愿意当受试对象。用图描绘此实验的一个区集设计。图 6.2 可当范本。
- (b) 用区集设计，比起拿 800 位受试对象来做完全随机化设计，有什么优点？区集设计比起用 800 位男性受试对象做完全随机化设计，又有什么优点？

6.17 比较减重计划。20 位体重过重的女性同意加入一项比较 A、B、C 及 D 四种减肥处理效果的研究。研究者先比较每个受试对象的实际体重和理想体重，计算出她超重多少。受试对象及其超重磅数如下：

伯恩鲍姆	35	赫南德兹	25	摩西	25	史密斯	29
布朗	34	杰克逊	33	尼夫斯基	39	史桃	33
布伦克	30	肯德尔	28	欧布拉	30	特兰	35
克鲁兹	34	洛伦	32	罗德里格斯	30	威兰斯基	42
邓恩	24	曼	28	圣地亚哥	27	威廉斯	22

反应变量是经过 8 周处理之后所减轻的重量。因为受试对象原本超重的磅数会影响反应，所以区集设计较合适。

- (a) 把受试对象依超重多寡排序，由小排到大。然后分成 5 个区集，每个区集有 4 个受试对象，分法是把超重的重量最小的 4 个放在一个区集，再来是次少的 4 个，以此类推。
- (b) 用表 A 来随机指派每个区集中的 4 个受试对象给 4 个减肥处理。要确实说明你是怎样用表 A 的。

6.18 玉米田。农学家想要比较 5 种不同品种玉米的收获量。实验准备用的那块田，肥沃程度是从北到南递增。因此农学家把田分割成



30 块一样大小的地，共有东西向的 6 列，每列再分 5 小块，然后用区集设计，以每一列为一区集。

(a) 画一个这块田的简图，分成 30 小块。把列分别编号为第一区集到第六区集。

(b) 执行区集设计的随机化，也就是说，把 5 种玉米 A、B、C、D、E 随机分配到每一个区集当中的 5 块地。在你的图上标示出来每一块地是种哪一种玉米。

6.19 加快邮递速度？信寄到另一个城市所需要的人数，跟几点寄出的，或是有没有写邮政编码有没有关系？简略描述一项实验设计来探讨这个问题，其中要有两个解释变量。处理是什么要确实说明清楚，还要说明你要怎样处理潜在变量，比如信是星期几寄的等等。

6.20 麦当劳对温蒂，在比较麦当劳和温蒂的吉士汉堡的盲目试验中（没有标示出哪个汉堡是哪家的），消费者会比较喜欢哪一家的产品呢？简略描述一下探讨这个问题的配对实验设计。

6.21 你想知道什么答案？前两题习题说明了怎样用统计实验设计来给日常问题找答案。选择一个你有兴趣，而又有可能利用实验来回答的问题，并大致讨论一项合适的实验设计。

6.22 医师与护士 美国的执业护理师(nurse practitioner)是有特定医疗护理专长的护士，他们通常就像提供初步保健护理的医生。一项实验把 1 316 位没有固定医疗资源的病人，有些指定给医师(510 位)，有些指定给执业护理师(806 位)。所有的病人在被指派之前，都已被诊断出有哮喘、糖尿病或高血压。反应变量包括 6 个月之后病人的健康评估，以及他们对所受医疗照顾的满意程度。

(a) 对病人的诊断(哮喘等)是处理变量(treatment variable)还是区集？为什么？

(b) 医疗种类(护士或医师)是处理变量还是区集？为什么？

第 7 章

数据伦理

假的手术是不是不道德？

“随机化、双盲、有安慰剂控制组的试验，是评估新事物的金科玉律，常用在评估新疗法上。”《新英格兰医学期刊》上一篇讨论帕金森氏症(Parkinson's disease)疗法的文章这样说。这篇文章并不是在讨论新疗法(虽然这个新疗法有希望可以减少疾病所带来的震颤和失控情况)，而是讨论研究这个新疗法时产生的伦理问题。

法律规定必须用设计完善的实验来证明新药有效并且安全。但是没有对手术做类似的规范，因此对于手术效果的研究中，只有 7% 做了随机化比较。外科医师都认为他们的手术成功，但是话说回来，所有创新的人都会认为自己的革新是成功的。即使病人病情真



有改善，说不定大部分的功劳该归给安慰剂效应。所以我们并不确知，许多常见手术是不是值得冒风险去做。要知道答案，得做合适的实验。合适的实验要包括扮演安慰剂角色的“假手术”。以帕金森氏症来说，看来颇有希望的新疗法牵涉到动手术植入新的细胞。安慰剂组也接受一样的手术，但是并不植入细胞。

这样子会不会不道德？有一边说：只有随机化比较实验，才能判别新疗法是不是有效。如果证实无效，可以免除成千上万病人的无谓手术；如果证实有效，那我们就可以对成千上万病人施行这个手术，而且知道效果不仅是安慰剂效应而已。另一边的论点是：任何手术都有风险。假手术让病人承受风险，却不能指望手术对他们有帮助，则医师不应该为了以后病人可能受益，而让目前某些病人承受风险。

对于利用假手术来检验帕金森氏症的新手术疗法是否有效，即使在很有理性的人之间也都有不同意见。但是对于统计研究的某些基本伦理原则，大家的看法一致。要考虑比较难解决的问题之前，我们应该先看看这些原则。

首要原则

产生数据和使用数据就像许多其他人类行为一样，都会引发伦理问题。例如一些推销员，明明是要用电话推销，却一开始就说：“我在做一项调查”，这种欺骗行为很明显是不道德的。而这样做也使合法的抽样机构非常生气，因为他们发觉一般大众比较不愿意接受他们的访谈了。还有少数的研究员，为了在专业领域更上一层楼，而发表不实资料。这不只是伦理问题，因为为了使你的事业有进展而伪造数据，根本就是错的，而一旦有人发现时，你的事业就完了。

至于对于真实没有做假的数据，研究员应该诚实到什么程度呢？下面的例子建议了答案：“绝对诚实，更甚于平时。”

最复杂的数据伦理问题产生于向人们搜集数据时。对人做某些治疗的实验和只搜集资料的抽样调查相比，前者牵涉到的伦理问题比较严重。



例 1 漏掉细节

科学研究报告应该简短且无赘言。但精简的结果常使研究员不把关于数据的事实全盘托出。选择研究对象的方式有没有偏差?是不是只报告了部分研究对象的数据?是不是试了好几种统计分析方法,然后只报告了看起来最好的结果?统计学家贝勒(John Bailer)于《新英格兰医学期刊》担任顾问的十年多间,曾审查超过四千篇医药论文。贝勒说:“从统计观点评论这些文章时,常常可以很明显看到,重要信息付之阙如,而漏失的部分几乎总是有一种很实际的效应,也就是使作者的结论看起来比实际该有的还要强。”在论文审查较松的领域,这种情况肯定更严重。

举例来说,新药的疗效试验可能对治疗对象有好处,但也可能有伤害。以下是数据伦理的一些基本标准,不论是抽样调查还是实验,任何要从人群中搜集数据的研究都应依照这个标准行事。

• 基本数据伦理

施行研究的机构必须设立**试验审查委员会**(institutional review board),负责事先审查所有的研究计划,以保护受试对象,使受试对象免于受到可能的伤害。

在搜集资料前,研究中的每一个受试对象都必须在**知情且同意**(informed consent)的情况下受试。

任何个人资料都必须**保密**(confidential),只有整体的统计结果可以公开。

美国的法律规定,任何由联邦政府补助的研究都必须遵照这些原则。可是不论是法律还是舆论专家,都不完全清楚全部的执行细节。

试验审查委员会

设置试验审查委员会的目的,并不是要决定某个试验研究计划是否能提供有价值的信息,或者计划的统计设计是否健全。委员会



的目的，套用某大学委员会的话：“是要保护被征召参加研究活动的受试对象(包括病人)的权利和福利。”委员会审核研究计划，而且可以要求做一些改变；委员会也审核同意书，以便确定受试对象会被告知研究的性质及可能的风险。一旦研究开始进行，委员会将一年至少监督一次进展情况。

审查委员会最迫切的危机是，他们的工作量过大，以至于使他们保护受试对象的效率降低。1999年当美国政府基于对受试对象的保护不够，而暂时中止杜克大学医学中心以人为受试对象的研究时，共有超过2 000件研究案例正在进行中。这可是任务很重的审查工作。

一些受试对象承受很少风险的计划，比如大部分的抽样调查，可以适用较短的审查步骤，当委员会工作量过大的时候，免不了会想把更多的计划归入低风险类，以加快审查速度。

知情且同意

在“知情且同意”这个语词当中，“知情”和“同意”两部分同样重要，而且也同样具有争议性。受试对象必须在事前被告知该研究的性质，及任何有可能发生的伤害。如果是抽样调查，当然不会对身体造成伤害，但受访对象应该被告知，调查中会问到哪类问题以及大概要花多少时间。而实验者必须告诉受试对象研究的本质及目的，并描述可能的风险，然后取得受试对象的书面同意。

即使对有行为能力的受试对象来说，“知情且同意”的困难仍然存在。有些研究者，特别是在医学实验中，认为要求同意是让病人参与研究的障碍。所以他们也许不会提到所有的风险；也可能不会说明有比正在研究的更好的疗法；即使同意书上有所有正确的细节，他们和病人说时也可能表现得过于乐观。从另一方面来看，提及每一项可能的风险，会使同意书又臭又长，而真的造成障碍。有个律师说：“同意书像租车合同一样。”有些受试对象看到5页或6页的同意书，根本就不读了。另有些受试对象被那么多可能发生(其实机会很小)的灾难吓坏了，结果临阵脱逃。当然，机会很小的灾难有时还是会发生。一旦发生，官司跑不掉，然后同意书就会变得更长也更详细了。



例 2 谁有能力同意?

是不是有些受试对象没有办法完全自主的“知情且同意”呢?举例来说,曾经有一段时间,新疫苗常常都以监狱囚犯做试验,囚犯如果同意,就可以得到行为优良记录做为回报。但现在我们担心,囚犯并不见得真的能依照自由意志拒绝,因此法律已禁止在监狱中做医药试验。

儿童也没能力做到完全知情然后同意,所以一般的程序是去问他们的父母。有一项对教导新阅读方法的研究即将在本地小学展开,因此研究团队也把同意书寄到家里给家长。不过很多家长都不把同意书寄回来,而因为家长没说“不行”,所以他们的小孩是不是可以参与研究?还是我们应该只让有家长寄回同意书并且同意的小孩参加?

对于精神病患者新疗法的研究要怎么办?救助急诊病人的新方法研究又要怎么处理?这些病人中有的陷入昏迷,或有的中了风。大部分的时候,连征求家属同意都来不及。“知情且同意”的原则是不是阻断了对昏迷病人实际试验新治疗方法的机会?

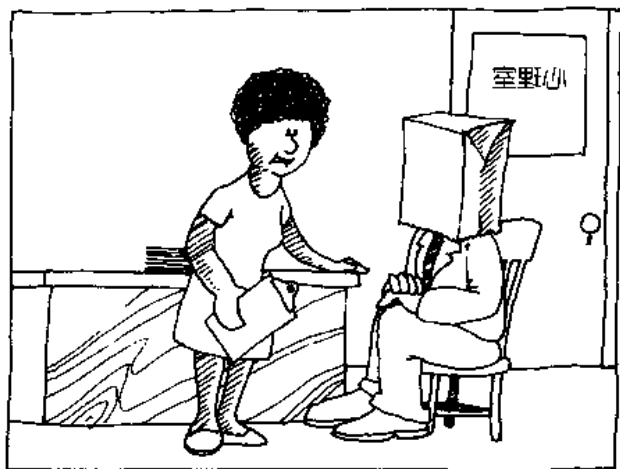
这些问题都没有清楚的答案。理性的人对这些问题也都有非常不同的意见。知情且同意可不是件简单的事。

保密原则

研究计划由试验审查委员会通过,并获受试对象同意,而且已取得关于受试对象的数据之后,伦理问题并没有就此消失了。保护受试对象的隐私权很重要,而要做到这点,必须对个人数据保密。民意调查的报告可以说,1 500 个应答者中,有多少比例认为美国的合法移民人数应该减低,但不可以说出“你”对这件议题或其他议题的意见。

保密和匿名(anonymity)不一样。匿名的意思是受试对象保持匿名身份,因此连研究主持人都不知道他们的姓名。匿名在统计研究中很少见。即使可能做到(主要是在邮寄问卷的情况下发生),匿名仍然有无法进行后续工作以增加应答率,以及无法将调查结果通知应答者的坏处。

任何违背保密原则的行为,都严重违反了伦理数据。最好的做法是,一开始就把受试者的身份资料和其他资料分开。比如说,在



“我知道这个研究是匿名参加的，但你还是得把眼睛露出来。”

抽样调查中，身份辨识资料只用来检查哪些人没有回答。然而，在现在这种高科技的时代，光是确知每一组资料都保障了隐私权还不够。比如说，美国政府保留了大量的公民资料，分别存在许多不同的数据库中，普查结果、所得税申报书、社会保险资料，以及类似当前人口调查之类的各种调查资料等等。这当中许多数据库的内容可通过电脑取得，以供统计研究用。即使你的名字和辨识身份的信息已由可供搜索的资料中去除了，只要用高超的电脑搜索技巧来搜索数个数据库，再把这些信息结合起来，仍有可能把你认出来，并得知许多关于你的信息。在这个电脑时代，数据的隐私和保密，在统计学家间已经是热门议题。

例 3 政府数据库的使用

公民有义务提供信息给政府，想想所得税申报及社会保险费就知道了。政府要这些数据是为了施政需要，也就是要看看我付税的金额对不对，以及到我退休的时候，政府欠我多大一笔社会保险给付。有些人认为，每个人应该有权禁止他的数据被用在任何其他方面，即使身份辨识信息已去除也一样。可是这样也同时禁止了利用政府记录来研究问题，比如获得社会保险补助者的年龄、收入和一户有多少人等问题。而在讨论如何改革社会保险制度时，这些研究很可能是必须的。



诚实和不诚实的 统计学家

发达国家依赖政府的统计学家来提供正确的数据。比如说,我们信任失业率的数字,也对公、私机构参考失业率所做的决定感到放心。但是可别理所当然地以为,全世界各地都是这样的情况。1998年的时候,俄罗斯政府逮捕了俄罗斯国家统计局委员会(State Committee for Statistics)中具领导地位的统计学家。他们被指控受贿制造数据帮公司逃税。某位报纸编辑说:

“这代表我们对俄罗斯的公司运营情况一无所知。”

临床试验

临床试验是为了研究疗效,而实际在病人身上进行的实验。医学疗法可能可以治病,但也可能对病人造成伤害,所以临床试验就成了以人当作受试对象的实验当中,道德问题的焦点。我们就从以下几点开始讨论:

- 随机化比较实验是可以肯定新疗法真正有效的惟一方法。若不做随机化比较实验,一些有风险且疗效顶多等于安慰剂的新疗法,就可能被普遍使用。
- 临床试验有很大的好处,然而大部分的好处是由以后的病人享用。试验也有风险,却由试验中的受试对象承受。所以我们应该在以后的好处和目前的风险之间找寻平衡点。
- 医学道德和国际人权标准都主张:“受试对象的利益,永远要摆在科学和社会利益之前。”

引号中的话引用自1964年世界医学会的赫尔辛基宣言(Helsinki Declaration of the World Medical Association),这是最受尊敬的国际标准。而不道德实验中最过分的例子,是不顾受试对象权益的那些。

因为“受试者的利益永远优先”,因此只有在有理由相信某项疗法,会对试验中的受试病人有帮助时,才可以进行临床试验来检验该疗法的疗效。对于以人当作受试对象的实验,光是以后对病人有好处不算充分理由。当然如果已经有强烈证据显示某个疗法有效而且安全的时候,不给病人用也不道德。哈佛医学院的亨内肯斯医师(Dr. Charles Hennekens)曾主持大型临床试验,证明阿司匹林能减低心脏病风险,他这样说过:

在何时应该做或者不该做随机化试验之间,存在着微妙的平衡。一方面必须对新疗法的可能疗效有足够的信心,才有理由让一半受试病人去尝试。另一方面又得对它的疗效有足够怀疑,才有理由让另外一半得到安慰剂的病人不试用这个疗法。

为什么给控制组病人安慰剂不算不道德呢?这是因为我们知道安慰剂通常对病人有帮助。除此之外,安慰剂也没有不良副作用。所以在亨内肯斯医师所描述平衡的怀疑情况下,安慰剂组得到的疗效,有可能比治疗组的还好。如果我们知道哪种疗法较好,我们就会让大家都接受这种



治疗。但当我们不知道的时候,两种都试并做比较,就没有不道德。

以下有一些和安慰剂有关,比较难回答的问题,并包含了正反意见。

例4 塔斯克吉梅毒研究

20世纪30年代,梅毒在美国南方乡下的黑人男性之间相当普遍,而这些人几乎没有什么医疗资源。公共卫生署(Public Health Service)从穷苦的黑人佃农中召集了399位梅毒患者和201位没有感染梅毒的人,来观察未经治疗时,梅毒病情会怎样发展。从1943年开始,已经可以用青霉素来治梅毒,但该研究的受试对象却没有获得治疗。事实上公共卫生署还阻止他们接受任何治疗,直到20世纪70年代消息走漏,研究被迫停止为止。

塔斯克吉梅毒研究(Tuskegee syphilis study)是研究者为了自己的利益而不顾受试者权益的一个极端例子。1996年的一篇评论中说:“这个事件代表了医学界的种族歧视,人类研究中的道德瑕疵,医师的父权心态,以及政府对弱势群体的滥用职权。”1997年克林顿总统在一项白宫仪式中,向幸存的受试者正式提出了道歉。

塔斯克吉研究事件也说明了,为什么今天有许多黑人由于不信任,而不愿参与临床试验。

例5 用安慰剂当控制?

你正在测试一种新药。如果有一种有效的药已经存在,还给控制组吃安慰剂是否合乎伦理原则?

是:安慰剂为新药的有效性提供了真正的基准。一共有3组来比较:新药、现存最好的药及安慰剂。每个临床试验都会有些不一样,即使真正有效的处理,也不见得在任何情况下都有效。用安慰剂当控制可以帮助我们审视:是否研究有瑕疵,使得现存最好的药连安慰剂都赢不了。有时安慰剂会赢,此时就有必要怀疑药剂的效用,因此绝对有理由使用安慰剂。所以除非是在性命攸关的情况下,否则用安慰剂当控制是合乎伦理的。



否：故意给病人比较差的治疗是不道德的。我们不知道新药是不是比现存的药好，所以两种药都给病人，以期找出答案，这是道德的。但是如果过去的试验已显示，现存的药优于安慰剂，再给病人安慰剂就不对了，毕竟现存的药也会有安慰剂效应。只有当现存的药是较久以前就有，因此未曾经过合适的临床试验，药效不理想或者有危险性的时候，给安慰剂才不会不道德。

例 6 假手术

安慰剂有效，而且当愈来愈多医师认清这个事实后，就有愈来愈多的医师有疑问：“如果我们在药品试验中可以接受安慰剂，为什么在手术试验中不也接受呢？”正如本章开头的帕金森氏症的例子中说明的，这是一个非常有争议性的问题。

同意：大部分的手术并没有经过比较实验的检验，其中有一些无疑只是安慰剂而已。然而跟安慰剂药品不一样，安慰剂手术会有风险。把真的手术和安慰剂手术做比较，可以免除掉数以千计的不必要手术，挽救许多性命。安慰剂手术可以做得很安全。比如说，可以给安慰剂组的病人一种安全的药，让他们不记得手术过程，而不必真的进行高危险性的麻醉。受试病人曾被告知他们参与的是一项有安慰剂控制组的试验，而他们仍同意参加。只要安慰剂组承受的风险很小，而且病人知情且同意，手术的安慰剂控制试验就合乎伦理原则（除非有性命攸关的情况）。

不同意：安慰剂手术和安慰剂药品不一样，总是有一些风险。记得前面说过的“受试者的利益永远优先”。除非这些受试者立即能享有某些好处，即使日后有很大的好处，也不足以成为让受试者承受风险的充分理由。我们可以给病人安慰剂药品当作治疗，因为安慰剂的确有效，而且没有风险。但因为手术有风险，所以没有医师会把它当做一般疗法。如果我们在治病时不会这么做，那么把它用在临床试验就是不道德的。



统计学上的争议

可以买希望？

我们已经指出以人做受试对象的实验，尤其是临床试验，有些什么样的伦理问题。但是没有做合适的实验也会造成问题。以下就是这样的例子。乳癌已到晚期的女性迟早会面临死亡，但此时出现了一种似乎有效，但尚未经过试验的疗法，我们现在就应该给病人试用的机会，还是等到有控制组的临床试验证明疗法有效之后再说？

看来很有希望的疗法是“骨髓移植”（BMT, bone marrow transplant）。BMT 的概念是先贮存病人的骨髓细胞，然后用高剂量的药杀死癌细胞，再把贮存的骨髓细胞放回病人体内，如此可避免好细胞和癌细胞同时被杀死。愈来愈多人用 BMT，即便这是很痛苦、很花钱又危险的疗法。

现在已有通过临床试验的抗癌药，但是对于 BMT 这类疗法，并没有禁制。在一些小型且没有控制组的 BMT 试验似乎成功之后，BMT 就被广泛使用了。而医学上的经济效益占了很大的原因。最早提供 BMT 疗法的是一些营利性质的医院，他们打很多广告来吸引病人，而其他人很快就跟进。《纽约时报》报道：“每一家提供这种实验疗法的机构，都用了不同的广告词——有些强调机构的声望，有些强调地理位置的便利，还有的则提及对病人的亲切照顾及支持……。”而这项

手术对医院和医师而言，利润都很高。

但是 BMT 比标准疗法更能延长病人寿命吗？我们并不确知，答案似乎是“不见得”。病人当然愿意尝试任何可能延长寿命的方法。有些医师则愿意提供没有足够证据支持的希望，病人不愿意参加控制试验，因为他们可能会被指派到传统疗法而非 BMT，由于召集受试对象不容易，要从这类试验得到结果会延误好些年。最先提出报告的 5 个试验当中，有 4 个结果并没有在 BMT 和传统疗法之间，找到有统计显著性的差异。第 5 个的结果则说是 BMT 疗效较佳，然而研究者很快就承认“严重违反了科学的诚实和正直原则”。《纽约时报》更直接地说：“他捏造数据。”

从情的观点出发，似乎大家同意濒临死亡的病人应可以使用尚未经过检验的疗法。但从理的角度来看，这样做会开启了贩卖希望之门，而且会对真正有效疗法的发展造成阻碍。来和儿童癌症的情况做个比较，对这些患儿，医师不同意使用未经控制试验的实验疗法。结果在所有癌症患儿中，有 60% 参加临床试验，而挽救患儿生命的进展，比挽救成年病人的进展要快得多。用 BMT 方法来治疗某种罕见儿童癌症，则立刻就进行了检验，并认定有效。难怪，最早用 BMT 治乳癌的开路先锋之一，在看到更多证据之后，现在承认：“我们欺骗了自己，也欺骗了病人。”



行为及社会科学实验

当我们从医疗问题转移到行为及社会科学时，受试对象的直接风险减低了，但是他们可能获得的利益也减少了。比如说，我们来看看心理学家为研究人类行为所做的实验。

例 7 别侵犯我的领空

心理学家观察到，每个人都有一个人“个人空间”，在别人太靠近的时候会不高兴。在咖啡店里的時候，如果还有别的空位，我们就不喜欢陌生人来和我们共桌，我们也看到人们在电梯里会尽量站开。美国人比大部分其他文化的人，更需要更多的个人空间。当个人空间被侵犯时，是否身体和心情都会受影响？

调查者在一个男公厕内动了手脚。他们把一些小便池封住，使得走进来尿尿的人，要么去用实验者(处理组)旁边的小便池，要么去用和实验者隔很远的小便池(控制组)。而另一个实验者在一个马桶间里，用潜望镜观察并度量受试者多久才开始尿尿以及尿多久。

这个个人空间实验说明了要计划和评估行为研究的人，面临了怎样的困难。

- 受试者并没有受到伤害的风险，虽然他们一定不同意别人用潜望镜偷窥他们。当身体受伤害的机会很小的时候，应该保护受试者哪方面？可能的情绪伤害？尊严受损？还是隐私权？
- 知情且同意又待如何？例 7 中的受试者根本都不知道自己参与了一项实验。许多行为实验都必须隐瞒真正的研究目的。如果事先知道调查者在观察什么，受试者的行为就会改变。受试者被要求同意时，根据的是模糊的信息。只有在实验结束之后，他们才知道全部的真相。



美国心理学会(American psychological Association)的“伦理原则”(ethical principle)要求,除非该项研究仅仅在公开场合观察行为,否则都要事先取得受试者同意。隐瞒只有在对研究必要时才可以,而且不可以隐瞒可能影响受试者参加意愿的信息,事后也要尽快地向受试者说明真相。个人空间研究(这是20世纪70年代做的)并不符合目前的道德规范。

我们已经见到,医学界和心理学界对于知情且同意的基本要求认知不同。底下的例子中有另一种情境,对于什么合乎道德又有另一种解释。受试者既不知情,也没同意;他们甚至浑然不知,有个实验正可能送他们去监狱过夜呢。



“我正在研究情绪紧张所产生的影响。现在,你过去拿我助手手上的斧头。”

例8 家庭暴力

警察接到家庭暴力报案的时候,应该怎么回应呢?在过去,通常的做法是叫施暴者离开,规定他整晚不准回家。因为受害者极少会提出控诉,所以警察并不愿意进行逮捕。女性团体主张,逮捕施暴者,即使不提出控诉,也会有助于防止暴力的再度发生。有没有证据可以证明,逮捕会减少暴力的发生呢?实验就是要回答这类问题。

一个典型的家庭暴力实验要比较两种处理:逮捕嫌犯并拘留过夜,或警告嫌犯后就放了他。当警察到达家庭暴力现场时,他们先让双方冷静后再调查。若有用凶器或威胁对方生命,就非得逮捕。若依情况可以逮捕,却不是非逮捕不可,警官就用无线电向总部请求指示。值班的人由文件夹中拿出最上面的信封,文件夹由统计学家事先准备好,信封内含处理方式,且顺序经过随机排列。警察根据信封的内容,逮捕嫌犯或警告后放了他。然后研究者会看警察的记录,并访问受害者,以了解有没有再发生家庭暴力。

这类实验第一次进行后似乎显示出,逮捕家庭暴力嫌犯,会减少他们以后的暴力行为。由于有这样的证据,逮捕已经成为警察对于家庭暴力的一般处置方式。



家庭暴力实验使我们对公共政策中的一个重要议题,有更多了解。因为没有受试对象的知情且同意,规范临床试验和大部分社会科学研究的道德条例会禁止这些实验的实施。但该研究被试验审查委员会批准了,因为,引用一位家庭暴力研究者所说的:“这些人是因为从事了让警察可以逮捕他们的行为,才变成受试对象的。你要逮捕一个人,可不需要经过他同意。”

网络寻奇

若想一窥试验审查委员会的工作内容,可造访一下美国匹兹堡大学(University of Pittsburgh)试验审查委员会的网站,网址为 <http://www.ihb.pitt.edu/>。看一下〈参考手册〉(Reference Manual,在该页第一项)就可以知道审查过程可以复杂到什么程度。

美国统计协会(American Statistical Association)和美国心理学会(American psychological Association)的正式伦理规范(网址分别为 www.amstat.org/profession/ethicalstatistics.html 及 www.apa.org/ethics/code.html)内容相当的长,除了本章讨论过的议题以外,还谈到许多其他议题。



本章重点摘要

根据一般诚信原则，你不应该捏造数据，或者实际是在卖盖屋顶的材料，却说是在做抽样调查。数据伦理的首要原则，不只是诚实而已。把人当做受试对象的研究，必须先经过**试验审查委员会**的审查。所有受试者在参与实验之前必须**知情且同意**。所有关于个别受试对象的信息应该予以**保密**。

这些原则是好的开始，但仍有许多争议存在，尤其是在以人为受试对象的实验当中。争议的内容都和受试者的利益及实验的未来利益之间的平衡有关。请记住对于某些问题，随机化比较实验是可以提供答案的唯一方式。还要记得：“受试者的利益，永远应该优先于科学及社会利益。”



第7章 习题

本章大部分的习题只是提出议题来讨论，所以答案无所谓对错，但是当然有是否经过周详考虑的差别。

7.1 最低风险？你是你的大学试验审查委员会的一员，你得决定有几件研究项目是否可以归类到低标审查类，因为那几个项目只会让受试者承受最低风险。美国联邦法规对于“最低风险”的解释，是该风险不会大于“日常生活中会遇到，或者在做例行身体检查或心理测验时产生的风险。”这种定义很模糊。你认为以下哪一项可以算是“最低风险”？

- (a) 在手指尖刺一下，抽一滴血来检查血糖。
- (b) 从手臂抽血，做整套验血程序。
- (c) 在手臂上固定一条插管，以便定时抽血。

7.2 试验审查委员会有什么人？美国政府规定，要求审查委员会要包含至少五个人，其中至少要有一位科学家，一位非科学家，以及一位机构外的人。大部分委员会不止五人，但其中有很多都只有一位机构外的人。

- (a) 为什么审查委员会里应该有非科学家？
- (b) 你感觉只一位机构以外的人够吗？你会怎样选择这位人士？（比如说，你会比较想选医师？神职人员？还是鼓吹病人权益的积极分子？）

7.3 试验审查委员会，你的大学里有一个试验审查委员会，负责审核所有把人当作受试对象的研究。去拿一份描述该委员会的文件的副本（大概可以上网取得）。

- (a) 根据这份文件，委员会的责任是什么？
- (b) 委员会的成员是怎样挑选的？有几位不是科学家？几位不在你的大学工作？成员有特殊专长，还是只是“一般社会大众”？

7.4 知情且同意。一位研究者怀疑，传统宗教信仰似乎和主张服从权力的个性有关。她准备了一份可度量服从权力倾向的问卷，并问了许多宗教问题。写下你对此研究目的的描述，这是要给受试



者看，以取得他们的知情同意书的。你得在两个互相冲突的目标之间求得平衡，一方面对于问卷会透露出受试者的何种信息不能有所欺瞒，另一方面也不能把信教的人给吓跑了，使得样本有偏差。

7.5 需要取得同意吗？在底下哪些情况下，你觉得可以不经受试者同意，就搜集个人资料？

- (a) 政府委托机构从所得税申报书中抽出随机样本，来取得有关不同行业的平均收入的资讯。他们不登记姓名，只记录收入和职业。
- (b) 社会心理学家参加了一个宗教团体的公开集会，来研究团体成员的行为模式。
- (c) 社会心理学家假装改变信仰，加入某个宗教团体，然后参加并不对外公开的集会，研究成员的行为模式。

7.6 以学生为受试对象。修心理学 001 课程的学生，被要求必须加入实验当受试对象。修心理学 002 的学生则并不一定要加入，但是如果加入实验，可以得到额外分数。修心理学 003 的学生可以二选一：不是选择当受试对象，就是要写期末报告。当实验中的受试者可能可以学到东西，但是以目前的伦理标准，对于用像囚犯或接受免费医疗的病人这类“不自主受试者” (dependent subject)，还是不能全无疑虑。学生相对于老师当然不算能完全自主。对以上几种课程安排你有没有反对意见？若有的话，是反对哪些安排，又为什么？

7.7 感染艾滋病毒的情况普遍吗？耶鲁大学的研究者和非洲坦桑尼亚的医学团队合作，想要知道在这个非洲国家的孕妇当中，感染艾滋病毒的情况有多普遍。为了要知道答案，他们计划要对孕妇抽血来检验。

耶鲁大学的试验审查委员会坚持，研究者必须取得每一位孕妇的知情同意书，并且告诉她们验血结果。这是发达国家的一般步骤。坦桑尼亚政府却不愿意告知那些孕妇为什么要抽血，也不愿意告知验血结果。政府担心如果验血结果发现许多人有这个不治之症，而国家医疗系统又不能提供治疗，会造成恐慌。最后研究被迫取消。耶鲁大学使用一般标准保护受试对象，你觉得做得对不对？

7.8 匿名还是保密？美国最重要的非政府调查之一是全面社会调查 (GSS，见第 1 章例 6)。GSS 定期检测一般大众对各式各样政治或社



会议题的意见，而访问是到受试者家中面对面进行的。受试者对 GSS 问题的回答是匿名的、保密的还是两者皆有？说明你的答案。

7.9 匿名还是保密？得克萨斯州农工大学(Texas A&M)和许多大学一样，提供免费的艾滋病毒(HIV)筛检，这种病毒会造成艾滋病。公告上说：“参加 HIV 筛检的人，会被分派到一个号码，所以可以不用报出姓名。”筛检的结果可以电话询问，同样还是不用报上姓名。这样做是匿名还是只是保密而已？

7.10 不是真的匿名，有些很普遍的做法看起来像匿名，但实际上只不过做到保密而已，市场调查者常常邮寄调查问卷，问卷上没有要回答者填姓名之类的识别资料，但是却有隐藏的编号，可以据以识别出回答者。如果宣称匿名却没匿名当然不道德。若只能做到保密，却不提有识别编号，使回答者可能误以为是在匿名回答，是不是也一样不道德？

7.11 人类生物物质，很久以前，医师在治你的轻微贫血时，曾经抽取你的血样。你有所不知的是，血样被储存起来。现在一些研究者计划要用你以及其他许多人被储存的血样，来找寻可能对贫血有影响的遗传因子，这当然不可能再来征求你的同意。以现今的科技，已经可以从你的血样读出你的整个基因组成。

- (a) 如果用有你名字的血样，但是当初却没告诉你血样会被留起来供以后的研究，你认为是否违反了知情同意原则？
- (b) 假如血样没有你的识别信息，只知血样属于(比如说)“一位 20 岁，接受贫血治疗的白种女性”。这样是不是就可以拿血样来研究了？
- (c) 也许对于血样这类生物物质，我们只应该使用同意让物质储存起来供日后研究用的病人的样本，但因为不可能先说明以后要做什么研究，所以这样还是没有达到一般的知情同意标准。但如果可以完全保密，而且使用血样不致对病人身体造成伤害的话，是不是还是可以接受？

7.12 一视同仁。老化问题的研究者要调查，额外的医疗服务对老人生活质量的影响。某家大型诊所病人名单上合乎条件的人，会被随机指派到处理组及控制组。处理组的病人会获得免费助听器、假牙、



交通以及其他服务，而控制组的人得自己付费。试验审查委员会认为，同一家诊所的病人，对一部分提供这些服务，其他的人却没有，也有伦理问题，你认为呢？

7.13 假手术？像本章一开始时提到的帕金森氏症研究这一类的临床试验，已经愈来愈普遍了。一位医学研究者说：“这只不过是刚开始而已。以后只要有新疗法，就得做双盲安慰剂试验。”例6当中大致说明了对于用检验药品的方式来检验手术的正反两面意见。什么样的情况下，你会同意在检验新手术的临床试验中执行假手术？

7.14 艾滋病临床试验。在终于有了艾滋病有效疗法的现在，再来检验也许疗效比较差的疗法，会不会不道德？同时使用几种强效药品的鸡尾酒疗法可以减低血液中 HIV 病毒的量，并且至少可以延缓艾滋病的发病及死亡。但是药的效果高低，还得看病人免疫系统的受损情况，以及他在之前用了些什么药，药的副作用很强，而且病人必须每天都能够准时服用总共超过 12 粒的药，因为艾滋病通常会致命，而且鸡尾酒疗法的药品组合很有效，我们可能认为，检验任何艾滋病的其他疗法，若是没有包含上述的完整药品组合，就是不道德的，但是这样可能对发现更好的疗法构成阻碍。这对于现在对病人提供已知最好的疗法，以及为未来的病人找出更好的疗法之间的冲突，提供了一个强有力的例子。我们怎样可以在合乎伦理规范的情形下，检验艾滋病的新药？

7.15 非洲的艾滋病试验。有效的艾滋病药品非常昂贵，所以大部分非洲国家负担不起给为数众多的病人这种药。然而艾滋病在非洲某些地区的盛行程度超过世界上任何其他地方。有好几个临床试验都在试着找寻防止受 HIV 感染的怀孕母亲把病毒传给胎儿的方法，而这种垂直感染是非洲 HIV 感染的主要来源。有些人认为这些试验不够道德，因为没有给受试对象治艾滋病的有效药品，而在富有的国家这是必须做到的。其他人却认为，试验目的是要找出在非洲真实世界中可行的疗法，而且这些试验至少会让受试者的儿女受益。你认为呢？

7.16 非洲的艾滋病试验。艾滋研究的最重要目标之一，是想找到可以预防感染 HIV 病毒的疫苗。因为艾滋病在非洲某些地区如此盛行，因此在这些地方试验疫苗会最容易。然而疫苗有可能太昂贵，以



至于无法(至少刚开始时)让非洲人普遍接种。如果在非洲做试验,好处却主要给富国享用,这样做道德吗?处理组的受试者会获得疫苗接种,如果证实疫苗有效,则安慰剂组之后也会获得接种。所以受试对象全部都会受益,而会丧失的是将来的利益。对此你有什么看法?

7.17 意见调查、美国总统选举战正在如火如荼进行,而候选人雇了民意调查机构定期做民意调查,以了解选民对一些议题的看法。做民意调查的人应该提供什么信息?

- (a) 依照知情同意标准,调查者必须告知可能的回应者哪些事?
- (b) 民意调查机构的共同规范中,也包括必须告诉回应者执行民意调查机构的名称和地址。你觉得为什么要有这项要求?
- (c) 民意调查机构通常有一个像“抽样公司”之类的正式名称,所以回应者并不知道调查是由某一政党或候选人付钱做的。如果告知回应者谁是出资单位,会使调查结果有偏差吗?是不是只要公布调查结果的时候,就应该一并公布是谁出的钱?

7.18 有权知道?有人认为应该立法规定,所有政治民意调查的结果都要公开。否则的话,拥有调查结果的人就可以利用那些信息来谋一己之利。他们可以根据信息采取行动,选择性地公开部分结果,或者选最有利的时机发布结果。有候选人的竞选委员会回答道,他们付钱做民意调查是为了取得信息给自己用,不是为了取悦社会大众。你赞成要求政治民意调查结果完全公开吗?其他私人调查,比如对消费者喜好做的市场调查呢?

7.19 回答政府问题。2 000 份繁式普查问卷上问了 53 个详细的问题,比如:

在你住的房子或公寓或活动房屋中,有完整的水管装置吗?这是指:(1)热水管及冷水管;(2)抽水马桶;(3)浴缸或淋浴装置。

问卷上还问了你的收入及来源,以及是否有“身体上、精神上或者情绪上的问题”造成你在“学习、记忆或专注上”的困难。国会中有些成员反对这些问题,虽然这个问卷是经过国会通过的。

分别对这个问卷的两旁不同论点做简短讨论:政府有权要求得到这类信息,但是这类问题似乎侵犯个人隐私。



7.20 使用数据要收钱吗?政府制作的数据,通常可以免费或低价供私人使用。举例来说,美国国家气象局(National Weather Service)的卫星气候资料,就免费提供给电视台在报告气象时使用,也免费提供给任何上网的人,而欧洲政府向一些国家的电视台提供气象资料时要收费。意见(1):政府资料应以最低价码供应任何人。意见(2):使用卫星很贵,电视台的气象报告又有收入,所以电视台应该分摊费用。

你支持哪个意见,为什么?

7.21 对年轻人做调查 疾病控制和预防中心(The Centers for Disease Control and Prevention)在一项对青少年的问卷中,问受试者是否有活跃的性生活。回答“是”的人即被问到下面这个问题:

你第一次性交时有多大?

问未成年人有关性、毒品使用以及其他这类问题时,需要取得父母的同意吗?或者本人同意就可以了?提出支持你意见的理由。

7.22 欺骗受试者,学生签名加入一项心理实验当受试者。他们到达时被告知,访问要稍延后,并被带到一间等候室。然后实验者安排让人偷走等候室里一样值钱的物件。有些受试者是单独和小偷在房间里,有些受试者是两人同时和小偷在房间里,而这就是实验者要比较的两种处理。受试者会不会告发这件失窃案?

学生先同意参加一项未指明的研究,实验的真正本质是在事后才向他们说明。你认为这项研究合乎伦理标准吗?

7.23 引诱受试对象。一位心理学家执行了下列实验:她先测量了受试者对作弊的看法,然后叫他们玩一项游戏,游戏动过手脚,使得不作弊要赢根本不可能。用来安排游戏的电脑也会记录下受试者有没有作弊,但受试者不知道这点。然后心理学家再次测试受试者对作弊的看法。

在游戏中作弊的受试者,倾向于改变他们的态度,对作弊比较能接受。而抗拒了作弊诱惑的人,在第二次测试时,倾向于更觉得作弊不对。这些结果证实了心理学家原先的理论。

这个实验引诱受试者作弊。并且误导受试者,让他们以为作弊神不知鬼不觉,而事实上是有人在观察的。从伦理观点应该反对这个实验吗?说明你的立场。



7.24 得体与否及公款 国会常常反对把公款花在看来似乎“不得体”或“没品味”的计划上面。最常受影响的是艺术，但是科学计划也可能遭殃。有一回国会拒绝了提供经费给一项研究大麻对性反应影响的实验。《科学》期刊的报导如下：

南伊利诺伊大学医学院(Southern Illinois Medical School)的鲁宾医师(Dr. Harris B. Rubin)和他的同行计划要给一些已吸食大麻的人看色情片，然后通过连在他们阴茎上的传感器来度量他们的反应。

大麻、性、三级片都到齐了，价格是120 000美元。但参议员用委婉的言辞以及各种隐喻，把这个烫手洋芋给“处决”了。

“我坚定认为，我们不缺赤字加下流电影加阿卡普尔金大麻*这样的组合，”阿肯色州参议员麦克莱伦(John McClellan)用强烈字眼表达了他的看法。

*译注：生长在墨西哥阿卡普尔地方的烈性大麻。

- (a) 受试对象是自愿参加的，而且在同意之前也得到了充分信息。如果你是试验审查委员会的一员，会不会根据是否得体或者有没有品味这类理由来否决这个实验？
- (b) 假设我们已同意，一个自由社会应该接纳任何有自愿受试者的合法实验。但是要认定任意这类实验，只要通过一般审查步骤认为它有科学价值，就应该得到政府补助，可是又往前迈了一步。如果你是国会一员，你会不会因为“不得体”或“没品味”这类理由，而反对补助一项实验？

第 8 章

度量

你有空闲时间吗？

现代人的空闲时间，比上一代的人多还是少？一本叫做《工作过度的美国人》（*The Overworked Americans*）的书当中说，我们比以前任何时候的人都劳碌得多。但另一本叫做《一辈子的时间》（*Time for Life*）的书，却说我们的空闲时间比已往任何时候都多。是不是有人用统计来说谎呢？

要知道空闲时间是否增加了，我们必须先要**度量**（measure）“空闲时间”。也就是说，我们必须把模糊的概念用会变化的数字表示。第一步先要表示清楚，我们说的空闲时间是什么意思？内容可能像“是指你不在工作，不在上、下班的路上，不在做家务，不在……的



时间”。你可以看出来,只要改变这一串“不在”的清单,我们得到的数字就不一样。

一旦我们决定什么是空闲时间之后,就得实际生产出这些数字,我们可以找几个人当样本,问他们昨天怎么过的。你不相信他们的记性?担心他们会夸大了工作时间?或许我们可以要求他们记下来,今天一天是怎么过的。当然真正很忙的人可能会忘了记下所有的工作时间。因此,不仅要明确定义“空闲时间”很难,而且无论我们怎么定义,要用数字去度量它也很难。

随机样本是很合理的,而随机化比较实验更合理。可是到头来我们还是得面对把“空闲时间”或“痛苦减轻了”或“收入”这些概念变成数字的问题。除非你知道这些数字怎么来的,否则不要轻易相信。

度量的基本原理

统计是讨论数字的。光是计划如何利用样本及实验来产生数据,并不会自动产生数字。一旦找到我们的回应者样本或实验受试对象,我们还必须度量我们感兴趣的特性。首先要大概考虑一下:我们准备度量的是不是正确的东西?有没有忽略什么也许不易度量却很重要的因素?

例1 病人怎样了?

临床试验倾向于度量容易度量的东西,像是血压、肿瘤大小、血液中的病毒含量。他们通常不直接度量对病人而言最重要的事情:经过治疗之后,病人的生活真的改善了吗?有一项研究发现,在1980—1997年之间发表的试验结果当中,只有5%度量了治疗对于病人在情绪方面和社交方面有什么影响。



一旦我们决定了要度量什么性质，就可以想想应该怎么度量。

量度

我们**度量**(measure, 也作量、评量或测量)人或物的某一性质，即是指用数字来代表那个性质。

通常我们用某种**器具**(instrument)来取得**量度**(measurement)。对于记录量度所用的单位(unit)，我们也许有不同选择。

量度的结果是一个**数值变量**(numerical variable)，不论我们量的是什么，只要我们要测量的人或东西在这项上有差别，这个变量的值就会不同。

例 2 长度、入大学适合性及公路的安全

我用卷尺当器具，来量我的床有多长。量度单位(unit of measurement)我可以选择英寸或厘米。若我选择厘米，则我的变量就是床的长度的厘米数。

要评量学生是否适合读大学，我也许叫学生去考“学术水平测验考试”(SAT, Scholastic Assessment Test)的推理部分。SAT 考试卷就是所使用的器具；变量是学生的考试分数，如果把 SAT 的语言部分和数学部分合计的话，分数大约会在 400—1 600 之间。

我要怎样度量公路上的行车安全呢？也许我会决定要用一年之中因车祸死亡的人数。政府的死亡事故通报系统(Fatal Accident Reporting System)搜集所有死亡交通事故的资料。我可以用政府资料中的交通死亡人数，当做度量公路安全的变量。

以下是你对任何统计研究中的变量应该问的问题：

1. 变量确实是如何定义的？
2. 该变量是否能有效描述它所声称要度量的性质？
3. 量度有多精确？

我们并不常自己设计度量方式，例如我们用 SAT 或者死亡事故通报系统的结果，所以我们不会深入探讨以上问题。但既然要用别人的数字，就得对它们有一些了解。



了解你的变量

量度是将诸如长度或受雇状况等概念,转换成明确意义的变量的过程。用卷尺可以直截了当地把“长度”这个概念变成数字,因为我们明确知道长度是什么意思。但要评量适合上大学的程度就有争议性,因为一个学生怎样才适合读大学并不完全清楚,用 SAT 分数至少明确说明了我们的数字会怎么来。要度量休闲时间,我们必须先说明,什么算是休闲时间。即使要算公路死亡人数,也得先说清楚怎样才算是公路死亡:被车子撞到的行人算吗?坐在汽车里而在平交道上被火车撞算不算?车祸 6 个月之后才因车祸中受的伤而死亡呢?没错,我们可以用政府的数字,但总有人得回答上面这些以及其他问题,才知道什么可以算进去。举例来说,一个人要在事故后 30 天之内死亡,才可算在交通事故死亡人数之内。这些细节挺烦人的,但是对数字有影响。

例 3 度量失业率

美国劳工统计局(BLS, Bureau of Labor Statistics)每个月都宣布上个月的失业率。没有准备要就业的人(比如说退休人士或者就学期间不想工作的学生),不能因为没有工作而算成失业人口。一个人要被归类为失业,必须先属于劳动人口;也就是说,他可以就业,也正在寻求工作机会。失业率为:

$$\text{失业率} = \frac{\text{失业人数}}{\text{属于劳动人口的人数}}$$

为了要把失业率确实实地定义清楚,劳工统计局对于“属于劳动人口”以及“失业”的定义,有非常详尽的描述。比如说你正在罢工,但是准备要回到原来的工作岗位,

你的单位是什么?

不注意量度的单位,会让你陷入大麻烦。1999 年的时候,火星气象观测轨道太空船在火星大气中焚毁。它原本应该处于火星上方 93 英里(150 千米),事实上则只到上方 35 英里(57 千米)处。这似乎是因为,该太空船的制造者,洛克希德马丁(Lockheed Martin)公司用英制(磅,英里)标示了重要量度,但负责太空船飞行的国家航空和航天管理局(NASA, National Aeronautics and Space Administration)的工作小组,却以为数字代表公制(千克,千米)。1.25 亿美元就这样泡汤了。



你就属于就业族。如果你没在工作，而过去两星期也没在找工作，你就不属于劳动人口。所以那些说想要工作，但是已经泄气而没有继续找工作的人，并不算是失业。细节是很重要的，如果政府用不一样的失业率定义，那么官方的失业率数字就会不一样了。



“这是我们新的入学标准测量器，这灵感是来自机场手提行李标准测量器。”

美国劳工统计局根据访问每个月“当前人口调查”的样本所得结果，估计出失业率。访问员不能只是问：“你属于劳动人口吗？”及“你有就业吗？”而是要问许多问题才能够将一个人归类为就业、失业或不属于劳动人口。改变问题可能改变失业率。1994年初，在规划若干年之后，劳工统计局开始采用电脑辅助访问并且将问题改进。次页的图8.1是失业率图，出现于劳工统计局有关就业状况的月刊头版上。因为访问过程做了改变，使得图在1994年1月的地方有个缺口。在旧制之下那个月的失业率应该是6.3%，照新制却是6.7%，这么大的差别让从政者很不高兴。

有效量度和无效量度

没有人会反对用以厘米为单位的卷尺，来量我的床有多长。却有很多人反对，用SAT分数当做是否适合读大学的指标。我们别争论这个，就只量量所有申请者的身高，然后录取个子最高的学生。傻主意，你一定会这么说。为什么呢？因为身高和适不适合读大学一点关系也没有。用比较正式一点的语言来说，身高并不是一个学生学业背景的有效(valid)量度。

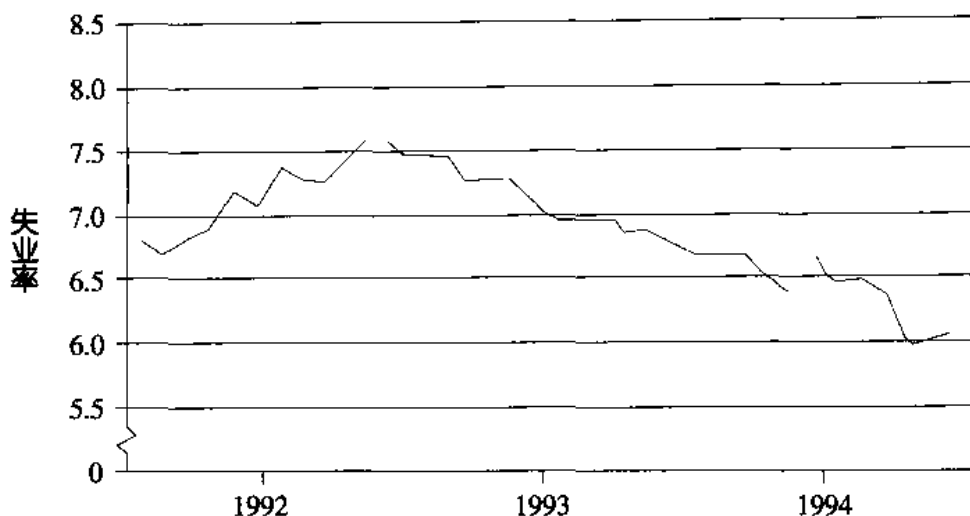


图 8.1 1991 年 8 月—1994 年 7 月的失业率。缺口显示出政府改变度量失业率的步骤所产生的影响

• 有效量度

当变数和某一性质有关，或者适合作为那个性质的代表时，我们称此变量为该性质之有效量度(valid measurement)。

用卷尺来量长度是有效的，而用学生的身高来评量她是不是适合读大学是无效的。劳工统计局的失业率是有效量度，即便改变度量失业率的定义会使量度稍有改变。让我们看看其他情境下的有效量度和无效量度。

例 4 评量公路安全性

路愈建愈好，速度限制增加了，大型多功能车取代了轿车，而且取缔行动减少了酒后驾驶。在这种变动的环境之下，公路安全有没有随着时间改变？

这只要看看车祸死亡数字就知道了。死亡事故通报系统说，1989 年有 45 582 人死亡，而 9 年之后的 1998 年，有 41 471 人死亡。但是有驾照的人从 1989 年的 1.66 亿，增加到了 1998 年的 1.85 亿。大家开车总共开的英里数，也从 2 096 000 000 000 增



加到 2 619 000 000 000。如果更多人开了更多英里,即使道路安全了,死亡人数也可能增加。死亡人数不是公路安全的有效量度。

因此用计数(count)来看道路安全并不理想,我们更应该用的是比率(rate)。计算每英里的死亡人数,就可以把如今更多人开更多英里的这个事实考虑进去。1989 年时,所有车子在美国总共开了 2 619 000 000 000 英里。因为这个数字太大了,所以评量安全性,通常是用每 1 亿英里的死亡人数,而不是用每 1 英里死亡人数来计算。以 1998 年来说,死亡率是:

$$\frac{\text{车祸死亡人数}}{\text{总开车的英里数(以 1 亿英里为单位)}} = \frac{41\,471}{26\,190} = 1.6$$

死亡率从 1989 年的每 1 亿英里 2.2 人,降到 1998 年的 1.6 人。这是很大的差距:1998 年和 10 年之前比起来,每英里死亡人数减少了 27%。开车已经愈来愈安全了。

• 比率和计数

通常来说,某件事情发生的比率(rate,或称百分比)和仅仅将发生次数做计数(count)二者比较起来,前者是较有效的量度。

用身高来评量适不适合读大学,或者在该用比率时却用计数,都是很清楚的无效量度的例子,而比较难搞定的问题,牵涉到既非很确定无效,也非明显有效的量度。

例 5 成果测验

当你在考统计这科时,你会希望考题涵盖课程大纲中的重点。这样的试题,是测量你对该课程题材有多了解的有效量度。主持 SAT 考试的大学委员会(College Board),也提供各种学科的学科测验(包括一个高等级的统计能力测验)。这些测验并没什么争议性。专家可以通过对比考题和考试范围的内容大纲,来评断考试的有效性。



例6 IQ 测验

心理学家想要测量人类性格中不能直接观察的一些层面，诸如“智力”或“权威性格”。IQ 测验能不能测量智力？有些心理学家会很大声回答“可以”。他们会说，有一种叫做“普通智力”的东西，而各种标准的 IQ 测验，虽然没法做到完美，但确实可以测量出普通智力。有些其他专家却说“不能”，而且声音一样大。他们认为，智力不是单一的，它由各种不同的心理能力共同组成，而没有任何单一器具可以量出各种不同的心理能力。

对于 IQ 测验是否有效有不同的意见，是植根于对于智力的本质看法不同。如果我们没有办法对智力到底是什么取得共识，也就没办法对该如何度量智力取得共识。

对这些例子中的问题，统计帮不上什么忙。问题的开始，是像“统计知识”或“智力”这样的概念，若概念本身就不明确，则有效性变成由个人主观决定。不过如果我们把有效性的概念变得更精确一些，则统计就很有用了。

例6 再谈 SAT

1999 年当 SAT 测验分数公布时，公正测验(Fair Test)机构表示：“SAT 测验的偏颇，会很不公平地让数以千计的年轻女学生丧失大学入学许可以及奖学金补助，而根据她们在学校里的优异表现，原应该得到这些的。”测验的数学部分性别差距比较大，女生平均 495，男生平均 531。联邦民权办公室(Federal Office of Civil Rights)说，女性和少数族裔考得比较差的测验，有歧视问题存在。

负责 SAT 考试事务的大学委员会回答，造成某些群体的平均分数比其他群体低的原因有很多。举例来说，来自低收入及低教育水准家庭而参加 SAT 考试的人，女生比男生多。平均来说，父母收入低且教育水平也低的学生，在家里和学校里的资源都不及有钱的同学多。他们的 SAT 分数比较低，是因为他们的背景使得他们为进入大学所做的准备不足。不能说他们分数较低就代表 SAT 不是有效量度。



量不出的仍然有关系

1981年石油人冰球队中的某位成员，差不多在任何可以度量的事项里都敬陪末座，包括：力量、速度、反应和眼力。那个人就是格雷茨基(Wayne Gretzky)，但很快的他就成为闻名的“天王”。他在那年打破了国家冰球联盟的得分纪录，接着在后来的7个赛季中得到更多的分数。不知怎的，一些很具体的量度都没能显示出格雷茨基是历史上最伟大的曲棍球员，所以并不是所有重要的特征都可以量出来的。

SAT是不是有效量度，可以检测有没有为读大学做好准备？“已经为读大学做好准备”是个模糊的概念，其中可能包含了先天的智力（不管我们怎么定义它）、学来的知识、读书方法、考试能力以及学习人文学科的动机。对于SAT分数（或任何其他量度）是否能正确反映这个模糊的概念，永远都会有不同的意见。

换个角度看，我们问一个比较简单，且容易回答的问题：SAT分数是否有助于预测学生能否胜任大学学业？能否胜任大学学业是很明确的概念，可以用学生能否毕业和他们的大学成绩来量。比起SAT分数低的学生，分数高的学生更有机会毕业，并得到（平均来说）较高的等级。我们说，以SAT分数做为是否准备好读大学的量度，有预测有效性。这是惟一可以用资料直接评估的有效性。

• 预测有效性

如果某一个性质的量度，可以用来预测跟这个性质有关的一些课题是否成功，我们称这个量度为**预测有效性**（predictive validity）。

从统计的观点来看，预测有效性是最明确而且最有用的一种有效性。“SAT分数是否有助于预测大学成绩？”这个问题，要比“IQ测验是不是可以测量智力？”明确多了。然而预测有效性可不是“是或否”的观念。我们得要问，用SAT测验来预测大学成绩的精确程度如何？还有我们得问，SAT是对怎样的群体有预测有效性？比如说有可能SAT预测男学生的大学成绩很准，对女性却不准。

有统计方法可以描述“精确程度”。在本章《统计学上的争议》中的讨论，就用了一种描述方法来说明整体状况。看来SAT分数的确有还不错的预测有效性，而且对不同群体的学生，有效程度也都差不多。不同群体学生SAT分数的差距，通常代表了他们有不同的环境，以至于对大学入学的准备程度也不一样。

准确和不准确量度

用家用体重计来量你的体重是有效的。可是如果你的体重计同我的体重计一样的话，量出来的体重就不见得很准了。来考虑一下我的



体重计，它可以量我的体重，但量出来的也许不是我的真正体重。我的体重计总是会多量 3 磅，所以：

$$\text{量出来的体重} = \text{真实体重} + 3 \text{ 磅}$$

如果事实果真如此，则对于同样的真实体重，体重计读出的数字就会相同。但是大部分体重计都会有少许变动，你离开体重计马上站回去，体重计的读数不见得会一样。我的体重计有点旧，不灵光了，因为没有真正归零再加上不稳定，它总是多量出 3 磅。今天早上机械有点不顺，使读数低了半磅。所以读数是：

$$\text{量出来的体重} = \text{真实体重} + 3 \text{ 磅} - 0.5 \text{ 磅}$$

我从体重计上下来再站上去的时候，体重计又往另一个地方停滞，使读数变成多 0.25 磅。现在我看到的读数是：

$$\text{量出来的体重} = \text{真实体重} + 3 \text{ 磅} + 0.25 \text{ 磅}$$

如果我闲着没事干，一直在体重计上上下下，就会前后得到各种不同的读数。读数会以超过真实体重 3 磅为中心，而上下变动。

我的体重计有两种误差。如果没有任何停滞情况，读数永远是多出 3 磅，不管谁站上磅秤都是这样。我们把这种每次度量时都发生的系统误差(systematic error)叫做偏差(bias)。我的体重计因为会停滞而不太顺，但是读数会改变多少，却是每次有人站上体重计时都不一样。有时停滞情况会使读数变高，有时却又使读数变低。结果就是，体重计平均来说会多秤出 3 磅，但是把同样一个东西重复过磅时，读数却会上下变动。这种因机械不顺而产生的误差我们根本无法预测，所以叫它做随机误差(random error)。

即便是有偏差，量同样的东西永远得到同样读数的磅秤仍是百分之百可靠的。而可靠的意思只是说结果会重复。偏差和不可靠(lack of reliability)是不同种类的误差。可是别因为可靠和有效都像是好的性质就把它们混淆了。即使体重计并不可靠，用体重计量体重仍是有效的。下面有某种量度可靠但却无效的例子。



统计学上的争议

SAT 测验及大学入学申请

美国的大学用各式各样的量度来决定收哪些学生。学生的在校成绩是最重要的，但 SAT 分数的影响也很大，尤其在某些很挑学生的学校。

SAT 的优点是，它是全国性的测验。在代数课中得到 A，在不同的中学有不同的意义，但是在 SAT 数学部分得到 625 分，代表的意义在全国各地都相同。SAT 没法量出一个人愿不愿意用功或者他的创造力如何，所以没法完全正确地预测进大学后的表现，但长久以来大部分大学都觉得它有用。

SAT 预测大一成绩的效果如何？本页最下方的表当中有一些结果，这是从 48 039 位学生的样本得来的。表里面的数字告诉我们，有多少百分比学生的大学成绩，可以用 SAT 分数（数学加上语言部分）、高中成绩或者 SAT 加上高中成绩来预测。数字若是 0 则代表没有预测有效性，而 100% 代表预测永远是完全正确。

我们可以看出用 SAT 预测大学成绩和用高中成绩预测的效果差不多。把 SAT 分数和高中成绩一起看，效果会比只看其中之一的要好。而对女生的预测事实上比男生的好一些。对黑人学生来说，预测的结果稍差些，但 SAT 的预测和高中成绩差不多。我们也看到，不论是 SAT 还是高中成绩，拿来预测大学表现都不算理想。高中成绩及 SAT 分数都差不多的学生，在大学里的表现却常常大相径庭。读书的动力和习惯对大学成绩的影响反而较大。

大学用 SAT 来当作是否收学生的参考是有理由的，但如果他们不只用 SAT 做决定，而且去看学生是否有强的学习动机，因为这动机能让底子不好的学生也可能成功，这样做也无可厚非。关于 SAT 的争论并不真的是对分数的争论。其实是关于大学应该如何运用所有的信息来决定谁可以入学，以及在决定谁入学时，大学的目标究竟是什么。

	所有学生(%)	男生(%)	女生(%)	黑人学生(%)
SAT	27	26	31	25
高中成绩	29	28	28	24
SAT 及高中成绩	37	36	38	34



• 度量时的误差

我们可以这样子看度量时产生的误差：度量出来的值 = 真正值 + 偏差 + 随机误差。度量过程如果有系统的量出比真正值大的值，或者有系统的量出比真正值小的值，就叫做有**偏差**。

度量过程如果在重复度量同一个体时，每次的值都不同，就叫做是有**随机误差**。若随机误差很小，我们称量度很**可靠**(reliable)。

例 8 大头里装的是聪明脑袋吗？

19 世纪中期，有人认为度量头颅的体积，就可以量出头颅的主人智力是多少，要可靠的度量头颅的体积很困难，即使当这个头颅已经和它的主人分家时，也不好度量。外科教授布洛卡(Paul Broca)指出，把头颅装满小型铅制弹丸，再把弹丸倒出来称重，可以相当可靠地度量出头颅的体积。但是这些准确的量度却不是智力的有效量度。头颅的体积事实上和智力或成就并没有相关性。

增加可靠程度，减少偏差

现在是什么时间？许多现代科技都要求能非常精确地度量时间，例子之一是全球定位系统(Global Positioning System)，它利用卫星讯号来告诉你，你的所在位置。时间根据地球绕太阳的路径起算，绕一圈是一年。但是地球太不稳定了，因此从 1967 年开始，时间根据标准秒来算，而标准秒的定义是铯原子震动 9 192 631 770 次所需要的时间。一般的钟会受温度、湿度和气压改变的影响，但铯原子却不理睬这些因素，所以需要非常精确计时的人可以买原子钟。美国国家标准与工艺研究所(NIST, National Institute of Standards and Technology)保存一个超级精确的原子钟，并且通过收音机、电话及互联网来报时(不过会因为传输时间，而产生一点误差)。



例 9 非常准确的时间

NIST 的原子钟非常准确，但并不是百分之百准确。世界标准时间是世界协调时 (Universal Coordinated Time)，是由位于法国塞夫尔的国际计量局 (BIPM, International Bureau of Weights and Measures) 所“编辑”的。BIPM 并没有比 NIST 更好的钟，它的时间，是用世界各地超过 200 个原子钟的时间平均得来的。NIST 告诉了我们(事后) 他们的时间距正确时间的差距。以下是作者正在写此书时 NIST 的最后 10 项误差，单位是秒：

0.000 000 007	0.000 000 000
0.000 000 005	-0.000 000 003
0.000 000 006	-0.000 000 005
0.000 000 000	-0.000 000 001
0.000 000 002	-0.000 000 001

长期来讲，NIST 对时间的量度并没有偏差。NIST 的秒有时比 BIPM 的短，有时比 BIPM 的长，并不是都较短或都较长。NIST 的量度很可靠，但是从上面的数字还是可以看出有些变异。世界上没有百分之百可靠的量度这回事。好几个量度的平均值，比起只量一次的结果，可靠程度比较高。这是 BIPM 要结合很多原子钟的时间的原因之一。

世界各地的科学家都利用重复度量，并且取其平均值来得到比较可靠的结果，即使学生在做化学实验时也常这样。就像比较大的样本可以减低样本统计量的变异一样，多用几个量度来平均，也可以减少结果的变异。

• 用平均来改进可靠程度

没有任何量度过程是百分之百可靠的。相较之下，重复度量同一个个体再取其平均，比单一量度更可靠(变异较小)。



不幸的是，却没有像这么直接的方法可以用来减低偏差。偏差大小是看度量器具有多好决定的。要降低偏差，你就需要好点的器具。NIST 的原子钟(图 8.2)的准确程度是每 600 万年误差 1 秒，但它要是放在你的床边，恐怕体积嫌大了些。

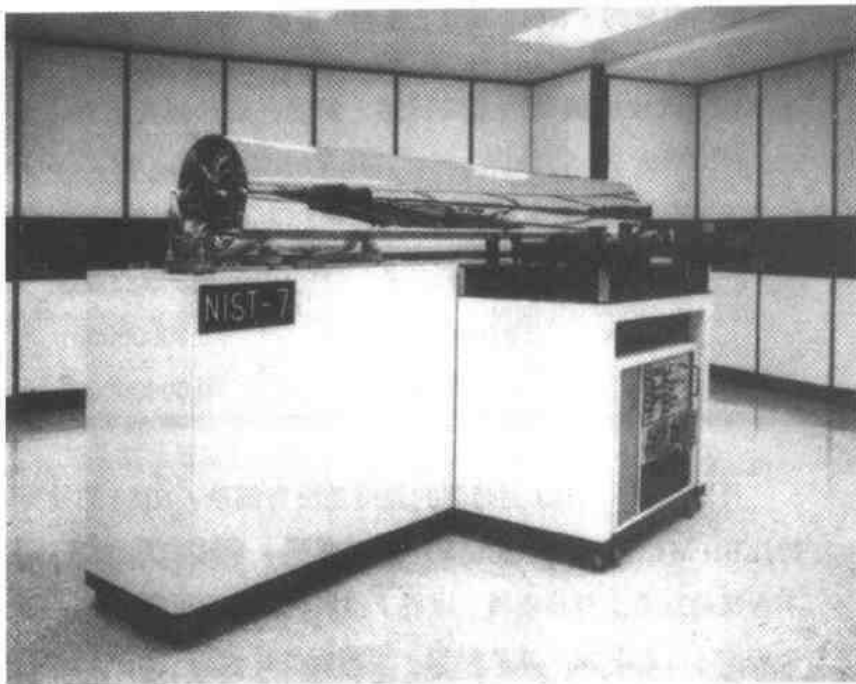


图 8.2 在 NIST 的这座原子钟，误差是每 600 万年 1 秒(Photo courtesy of John Wessels, NIST Time and Frequency Division)

例 10 再次度量失业率

度量失业率也是“量度”。就如同度量时间一样，对度量失业率而言，偏差和可靠程度的观念同样适用。

劳工统计局会叫监督员重新访问 5% 的样本，来检查他们所度量的失业率是否可靠。这就是对同一个体的重复量度，就像学生在化学实验室里，量一个东西的重量好几次一样。

劳工统计局通过改善器具来减少偏差。1994 年发生的就是这回事，那时当前人口调查进行了超过 50 年以来的最大翻修。比如说，度量失业率的旧系统把女性失业率低估了，因为度量的细节过程，没有随女性工作形态转变而跟着调整。新的度量系统更正了这个偏差，因而提高了所报道的失业率。



请同情可怜的心理学家

统计学家是习惯的动物：他们对于量度的考虑，就和他们考虑抽样时差不多。两种情况下的主要概念都是要问：“如果我们重复做很多次，会发生什么状况？”在抽样时我们想要估计一个总体参数，得担心估计值是不是有偏差，或者不同样本之间估计值的差别是不是太大。现在我们想要度量某个性质的真正值，就担心我们的量度也许有偏差，或者当我们重复度量同一个体时变异太大。偏差是每一次都发生的系统误差；高变异（低可靠程度）表示我们的结果因为不具重复性（not repeatable），所以不可信任。

当我们在量我的体重的时候，这样子考虑量度是很直接的方式。首先我们对“真正重量”代表什么有很清楚的概念。我们也知道世界上有很好的磅秤存在：我可以先考虑去诊所量，或者去物理实验室量，甚至到 NIST 去量。我们可以把我的体重量到我们想要的精确程度。这样很容易可以发现，我家的磅秤，总是把我多秤了 3 磅。而可靠程度也很容易描述：在体重计上上下下很多次，看看读数变化有多大。

然而当我们想度量“智力”或“是否适合读大学”的时候，若要知道“如果我们重复许多次，会发生什么状况”的话，执行起来可要困难得多了。我们来看看，可怜心理学家想要度量“权威人格”（authoritarian personality）的例子。

例 11 权威人格？

是不是有些人的性格，使他们想法比较僵硬而且会追随有力的领导者？在二次大战之后回头研究纳粹党员的心理学家认为，的确如此。1950 年的时候，一群心理学家发明出“F 量表”当做度量“权威人格”的器具。F 量表询问你，对于类似以下叙述在多大程度上同意或不同意。

- 服从和尊敬权威是儿童最应该学习的美德。
- 科学的地位不容怀疑，然而世上还是有许多重要的事，人永远也不会懂。



你要是非常同意这类叙述，就会被视为有信服权威的性格。F量表以及权威人格这个概念，在心理学中一直都很重要，尤其是在研究偏见和右翼极端分子活动的时候。

对于用F量表来度量“权威人格”，我们也许会问下面这些问题。当我们在考虑IQ测验或者SAT测验的时候，同样的问题也会浮现在脑海。

1. 到底什么是“权威人格”？我们对这个概念的了解程度，还不如我们对我体重的了解。实际的回答似乎是“反正就是F量表在度量的东西”。要说这种量度有效，必须先知道F量表的高分数代表什么样的行为，也就是说我们仍在考虑预测有效性。
2. 权威人格听起来不好听，而且F是代表法西斯主义者(Fascist)。就如例11中第2个问题所暗示的，有传统宗教信仰的人和非常类似但却没有宗教信仰的人比起来，前者较可能在F量表中得高分。度量器具是不是反映了发明出这套器具的人自己的想法，因而换个不同想法的人就可能发明出很不一样的器具呢？
3. 我们自认为了解自己真正的体重是多少，那我真正的F量表分数应该是多少呢？NIST可以帮我们找出真正的体重，但是没法帮我们找到真正的权威人格的分数。如果我们怀疑度量“权威人格”的器具具有偏差，因为它对有宗教信仰的人不公平，我们要怎么样去检测呢？
4. 我们可以把我秤很多次，来了解我家体重计的可靠程度。如果我接受F量表的测验很多次，我会记得我第一次写了些什么答案。也就是说，重复同样的心理量度很多次，并不能算是真正的重复。所以实际上很难检测出可靠程度。也许心理学家可以把同一项器具，发展出好几个不同的形式来执行重复量度。但是我们又怎样能知道，这些不同形式的器具效果是不是真的一样呢？



重点并不是说心理学家对以上问题都提不出答案。头两个问题本来就有争议性，因为并不是所有心理学家对人类性格的思维路线都一样。后两个问题至少有部分答案，但是答案并不简单。问题是，在我们度量体重时，“量度”这个词的意义十分清楚明了，但在我们想要度量人类性格的时候，可就变得极其复杂了。

这里还有个更重要的课题。当你看到诸如信服权威性格、智力甚至入大学适合性这类不明确主题的相关统计“事实”时，一定要小心。数字看起来总是很可靠。但是数据是人制作出来的，因此会反映出人的欲望、偏见和弱点。如果我们对到底在度量什么都不了解也未取得共识，则数字可能会制造争议而不是澄清问题。

网络寻奇

你可以从 www.time.gov 直接得到 NIST 原子钟的时间。虽然会因为互联网的耽搁而有一些误差，但网页甚至还告诉你，屏幕上看到的时间大约有多精确。若想一窥我们日常度量的长度、重量和时间背后的复杂系统，可以访问国际计量局的网站：
www.bipm.fr。



本章重点摘要

度量一个东西的意思，是把一个个体的某一性质用数字来表示。当我们度量很多个体的同一性质时，就得到同一个**变量**的许多不同值，而这个变量就是用来描述该性质的。当你处理数据或者读到统计研究的结果时，要弄清楚变量的确实定义，以及他们是否漏掉了些你想知道的事情。要确认研究中的变量是否是所讨论的概念之**有效量度**。

对于物理性质的量度，例如长度、重量及时间，有效与否很容易判断。当我们要度量人的性格或其他模糊性质时，**预测有效性**是用来判断“我们的量度是否有效”最有用的方法。此外还应该了解一下在取得数据时，是否有**量度误差**(error in measurement)，以致降低了数据的价值。量度误差可以看成下面这样的情况：

度量出来的值 = 真正值 + 偏差 + 随机误差

有些度量方法是有**偏的**，即有系统的偏向同一方向。要减低偏差，你须用好一点的**器具**来度量。有些度量过程又不够**可靠**，因此重复度量同一个体，会因**随机误差**而得到颇不一样的结果。要增加一个量度的可靠程度，可以多量几次再取平均。



第8章 习题

8.1 要算失业人数吗?要知道失业情况可以计数(看有多少人失业),也可以算比率(属于劳动人口的人当中,失业者所占百分比了)。在美国属于劳动人口的人数,从1980年的1.07亿,增长到1990年的1.26亿,再增长到2000年初的1.4亿。利用这些事实来说明,为什么失业人数不是失业状况的有效量度。

8.2 度量健康状况 你想度量大学生的“健康状况”(physical fitness)。举个例子说明度量健康状况明显无效的一个方法,然后简略描述一个你认为有效的度量过程。

8.3 校车安全 美国国家公路交通安全管理局(National Highway Traffic Safety Administration)宣布,每年平均有11个学龄儿童死于校车车祸,而有平均600个学龄儿童在上课时段死于一般车祸。从这些数字看来,似乎搭校车上学比乘坐家长开的车上学要安全。但是光看这些计数还不够。要比较校车和私家车何者较安全,你觉得应该用什么比率来比才好?

8.4 比率及计数 在某假期中,顾客退了36件大衣给西尔斯百货公司,而隔壁的精品服装店只有12件大衣被退。西尔斯本季卖出1100件大衣,精品店本季卖出200件。

(a) 西尔斯被退的大衣件数比较多。为什么不能据此判断,西尔斯的大衣顾客和精品店的比起来,前者较不满意?

(b) 两家店的退货率(被退回大衣所占百分比)各是多少?

8.5 死刑 在1977—1998年之间,美国有432名已定罪的囚犯被执行死刑。以下是7个州在那段时间中执行死刑的人数,以及那些州在1990年的人口数:

州	人口数(单位:千人)	执行死刑人数
亚拉巴马	4 040	16
阿肯色	2 351	16
佛罗里达	12 938	39



密苏里	5 117	29
内华达	1 202	6
得克萨斯	16 986	144
弗吉尼亚	6 189	46

得克萨斯州和佛罗里达州在处决人犯的人数上都名列前茅。因为这两州是大州，所以执行较多死刑也在意料之中。算出表中7个州当中每一州的处决比率，以每百万人口的处决数来计。因为人口数是以「人」为单位，所以每百万人的比率可以用下面公式来算：

$$\text{每百万人比率} = \frac{\text{处决人数}}{\text{人口数(以千人为单位)}} \times 1\,000$$

根据这个相对于人口数的处决人数的多少，重新把7个州排序。佛罗里达和得克萨斯是否仍然名列前茅？

8.6 度量智力 “智力”的意思，差不多是“一般的解决问题能力”。解释一下为什么问以下这类问题不能有效度量智力：

美国国歌的歌词是谁写的？

哪一队赢了上次的世界杯足球赛？

8.7 度量生活质量 英国人的生活是愈过愈好，还是愈过愈差？从一般政府资料无法反映出来，所以英国政府宣布，准备开始度量诸如居住、交通及空气污染状况。一位副首相说：“生活质量不光指经济方面而已。”帮帮他们的忙：你觉得“交通状况”和它对生活品质的影响要怎么度量？

8.8 度量疼痛程度 美国退役军人事务部替340万病人提供医疗服务。该部希望医师和护士把疼痛当做“第五种生命体征”，要和血压、脉搏、体温及呼吸一起记录下来。帮帮他们的忙：病人的疼痛你要怎样度量？

8.9 对抗癌症 美国国会要求医疗单位提出抗癌有进展的证据。可以考虑用的变量有：



- (a) 因癌症死亡的人数。这个数字随时间大幅上升,从1970年的331 000,到1990年的505 000甚至到1998年的539 000。
- (b) 美国人死于癌症的百分比。死于癌症占全部死亡的百分比,从1970年的17.2%稳定增加到1990年的23.5%,然后在1998年的23.0%稳住。
- (c) 从发现疾病时算起,存活超过5年的病人比率。这个比率在缓慢上升中。对白种人来说,5年存活率在1974—1976年间是50.3%,1989—1995年间是60.9%。

作为治疗癌症有效性的量度,以上这些变量没有一个是完全有效的。解释一下为什么即使治疗愈来愈有效,(a)和(b)仍然可能增加;而即使治疗愈来愈没效,(c)还是可能增加。

8.10 测验求职者。美国法律规定,对申请工作者做的测验,一定要和工作直接相关。美国劳工部认为有一项叫做“一般能力性测验”(GATB, General Aptitude Test Battery)的招聘考试,对于很大范围的各种工作都适用。就像SAT测验一样,黑人和西班牙语系的人,平均GATB成绩要低于白人。简短描述要怎样做才能说明,以GATB当做未来工作表现的量度,有预测有效性。

8.11 有效性、偏差及可靠程度 请举一个有效然而偏差人的度量过程的例子。然后再举一个度量过程的例子,是无效然而很可靠的。

8.12 偏差实例。让我们来看看直觉量度的偏差如何。图8.3中画了一个倾斜的杯子,影印或照着画10张。找5男5女共10个人,向他们说明这个图代表一个倾斜的水杯。要求每个人画出装满水时的水面。

正确的水面是水平的(和杯口最低处齐)。很多人在估计水面时会有不小的误差。用量角器量出每个人的角度误差。这些人的误差是不是都朝同一个方向?平均误差有多大?男性和女性的平均误差有没有明显差别?

8.13 偏差及可靠程度之实例。准备5段绳子,长度分别为2.9、9.5、5.7、4.2及7.6英寸。

- (a) 把这几段绳子一段一段地拿给一个学生看,要求学生靠目测来估计每段绳子的长度,估计到英寸的小数第一位。你的受试者的估

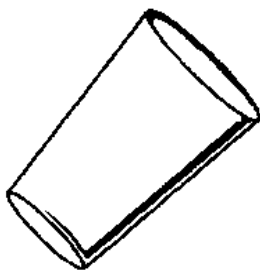


图 8.3 习题 8.12 描述的倾斜的杯子。如果杯子装满水，你画得出水面的位置吗？

计误差，是他目视得到的值减掉真正的值，结果可能是正的也可能是负的。5 个误差的平均值是多少？解释一下，如果没有偏差，而且我们用很多段绳子而不是只用 5 段的话，为什么平均误差应该靠近 0。

- (b) 第二天叫同一位学生再估计一次每段绳子的长度（把顺序换一换）。如果这位受试者度量的长度完全可靠，说明一下为什么每段绳子，第一次猜和第二次猜的长度差距都是 0。差距愈大，就表示受试者愈不可靠。你的受试者的平均差距（算平均时不计入正负符号）是多少？

8.14 续论偏差及可靠程度 上一题里有 5 段绳子的确实长度。一个受试者靠目测把每段绳子量了两次。依 (a) 和 (b) 的描述，分别虚构出符合描述的目测结果。为了简单起见，就把偏差当做是每次量都产生的同样误差，而不是量很多次的——种“平均”误差。

- (a) 受试者的偏差是过长 0.5 英寸，而可靠程度是百分之百。
(b) 受试者没有偏差，但可靠程度不是百分之百，重复度量的平均差距是 0.5 英寸。

8.15 职业训练有用吗？要度量政府的职训计划是否有效，常常会比较参加职训的人在训练之前和之后的收入差距。但是很多人会参加职训，就是因为收入减少或者被裁员了，所以“之前”的收入通常很低，使得收入的增加幅度显得很大。

- (a) 在度量职训对收入的影响时，以上现象属于偏差还是随机误差？为什么？
(b) 你会怎样度量职训计划是否成功？

8.16 怎样会造就不可靠的结果 政府每个月发表“国民储蓄”的



资料，这个数字告诉我们，大家在上个月存了多少钱。储蓄金额这个数字，是用大家的个人收入(很大的一个数字)，减掉大家的个人支出(又一个很大的数字)得来的，所得结果就是政府最不可靠的统计资料之一。

用一个数字的例子来说明，两个很大的数字分别只有小百分比的改变，也可以造成这两个数字差距的大百分比的改变。一个变量如果是两个很大数字的差，通常不会太可靠。

8.17 评量犯罪情况。犯罪资料常常登在报纸的头条。我们评量犯罪情况，可以用犯罪的件数，或者(较好的选择)用犯罪率(每 10 万人口的犯罪件数)。FBI 整合各警察局的报案纪录，并公布全美国的犯罪资料。美国全国犯罪案件受害者调查(National Crime Victimization Survey)公布的，是根据 43 000 户住户的全国概率样本所得到的资料。受害者调查显示的犯罪件数，是 FBI 报告的大约两倍半。说明一下为什么 FBI 报告，对于多种型态的罪案件数都会出现大幅偏低现象。(这是产生数据时的偏差导致量度误差的一个例子。)

8.18 评价犯罪情况。美国全国罪犯案件受害者调查每年都会询问至少 43 000 个住户的随机样本，他们是否曾是罪犯案件的受害者，若答是的话，他们还会问细节。总共有超过 8 万人回答这些问题。如果当一个人在回答问题的时候，同住户有别人也在同一个房间中，那么例如像强暴或其他性侵害案件的量度，就可能有很大偏差。为什么呢?有其他人在场，会使得性侵害案件多报还是少报?

8.19 量脉搏。你想量你自己静止时的脉搏。你可以量 6 秒钟的搏动次数，然后乘以 10 来得到每分钟搏动次数。为什么这样量，不如实际量 1 分钟的次数来得可靠?

8.20 你读的大学优秀吗?《美国新闻与世界报道》(*U. S. News and World Report*)每年都发表全美大专院校的排名。你可以在该杂志的网址上看到最新排名，网址是 www.usnews.com。许多大学常常对于排序的有效性提出质疑，有些甚至说，美国新闻每年都把度量标准稍微更动，以便产生不同的“赢家”好制造新闻。该杂志在它的网站上描述了他们的排序方法。提出三个你会用来当做大学“学术卓越”指标的变量，不一定要在美国新闻的变量里挑选。



8.21 很难通过的课程? 一个朋友告诉你:“美国历史这门课有 20 个学生未能通过,而俄国历史只有 11 个学生未能通过。所以美国历史的教授打分数要比俄国历史的教授严格。”说明为什么他的结论未必正确。要比较这两班你还需要什么信息?

8.22 测验求职者 一家公司以前对所有求职者都测验 IQ,但现在这样做变成违法,因为并不是这家公司里的所有工作的表现好坏,都和该项工作操作者的 IQ 有关。这项政策改变的原因所牵涉到的,是以 IQ 测验做为未来工作表现量度的可靠程度、偏差还是有效性?试说明之。

8.23 较好的洗碗机 你在为一本消费者杂志写一篇文章,内容是根据向杂志读者询问他们的家电是否可靠的一项调查结果。在回答拥有 A 牌洗碗机的 13 376 位读者中,有 2 942 位在去年一年中需要报修;而拥有 B 牌洗碗机的 480 位读者中,只有 192 位报修。提出一个合适用来评估不同品牌洗碗机可靠程度的变量,并对 A 牌和 B 牌洗碗机分别计算这个变量的值。哪个牌子比较可靠?

8.24 住哪最好? 每年《金钱》(Money)杂志都会在一篇谈论哪里最适合居住的文章中,替全美 300 个大都会区排名。1997 年的第一名是罕布什尔州的纳舒厄。纳舒厄在 1996 年排名 42,1995 年排名 19。新泽西州的蒙默思及海洋郡在 1995 年排 167 名,1996 年 38 名,1997 年第 3 名。这些证据是显示《金钱》杂志的排名方法无效、有偏还是不可靠?说明你的选择。

第 9 章

数字合不合理？

消失的厢型车

通常汽车制造商会借钱给他们的车商，好让车商有车可以展销，等到车子卖出时，车商再偿还贷款。纽约长岛有一位叫做麦纳玛拉的车商，在 1985—1991 年之间，向通用汽车公司借贷了超过 60 亿美元。光是在 1990 年 12 月，麦纳玛拉先生就借了 4.25 亿，购买了 17 000 辆通用厢型车，交由印第安纳州一家公司来改装，号称要销售到海外。因为麦纳玛拉信用良好，所以通用汽车欢欢喜喜地借钱给他。

让我们暂停一下，仔细想想这些数字，这是通用汽车该做却没做的。整个厢型车改装的行业，一个月差不多只生产 17 000 辆厢型



车,所以麦纳玛拉等于宣称他一个人买下了全美整个月的生产量。而这些豪华又费油的大型车,是为了美国的州际公路设计的,休闲旅行车贸易协会说,这种车在1990年只有1.35%的车辆外销。宣称在一个月內要买17 000辆厢型车来外销,是令人怀疑的。即使和厢型车的全部生产量相比,麦纳玛拉所宣称购买的量也是够大的。以雪佛兰公司为例,1990年整年才生产了100 167辆正常大小的厢型车。

看到这些数字之后,你也猜得出是怎么回事了。麦纳玛拉于1992年在联邦法庭中承认,他大大地诈骗了通用汽车公司。印第安纳的那家公司是麦纳玛拉建立的空壳公司,它的发票是伪造的,号称购买的厢型车根本不存在。麦纳玛拉大量向通用汽车借钱,其中大部分用来还前次的借款(因此而建立了良好的信用纪录),而自己也揩了一些油。前后加起来,他总共诈骗了4亿美金。通用汽车拿出2.75亿来补这个漏洞,而有两位应该要仔细审查相关业务数字的主管,也因此被开除了。

商业数据、广告主张、对公共议题的辩论——我们每天都被数字围攻,这些数字的目的是要证明观点、加强论据或者只是要我们放心。有些人比如像麦纳玛拉,给我们假数据。其他人用数据来为某个目标争论,但是对目标的关注,甚于数字的正确性。还有人简直没有处理数字的能力。我们知道永远要问:

- 数据是怎样产生的?
- 所度量的事实是什么?

我们也相当了解,怎么样的答案才像是这些问题的好答案。这样子很棒,但是还不够,被麦纳玛拉骗了的通用汽车主管,有可能知道随机样本和可靠量度,他们缺的是“数字感”,就是检验数字是否合理的习惯。为了帮大家建立起数字感,我们来看看,坏数据或者错误使用的好数据,怎样可以让粗心的人上当。



他们在说什么？

最常见的误用数据方式是，虽然引用了正确的数字，但因为没把事实全说出来，所以数字的意义并非表面上看起来的那样。数字并不是捏造的，所以信息有点不完整也许只是无心的疏失。这儿有些例子，你自己决定他们到底有多“无心”。

例 1 雪！雪！雪！

Crested Butte 打广告说，在所有科罗拉多州的滑雪镇里，它的平均降雪量最高，借此吸引滑雪者。这是事实。但滑雪者需要的是雪下在滑雪坡，不是下在镇上，而在许多其他的科罗拉多度假胜地，滑雪坡上的雪比较多。

例 2 又一个雪的例子

对于暴风雪，新闻报道会这样说：“这场冬季风暴挟着雪横扫该地区，造成 28 起轻微交通事故。”威斯康星州密耳瓦基的记者迈尔说，他常打电话给警长，以搜集这类资讯。有一天他决定要问警长，好天气的时候通常有几起轻微交通事故？警长说大约 48 起。迈尔说，也许新闻中应该这样写：“今天的暴风雪，防止了 20 起轻微交通事故的发生。”

这些例子的重点是，数字是有相关内容的。如果你不知道相关内容，那么单独、赤裸裸的数字就没法给你太多信息。



例3 我们吸引优秀的学生

各大学都知道，许多学生都会参考某些很受欢迎的入学指南，来决定要向哪个学校申请入学。指南中所印的信息是大学自己提供的，当然没有哪个大学会对一些资料，比如说该校入学生的平均 SAT 分数，公然说谎。可是，我们确实希望分数看起来很好，是不是可以不要计入外国学生以及需要补习的学生的分数？波士顿东北大学(Northeastern University)就这么做了，结果使得新生平均 SAT 分数，比起包括所有学生算出来的分数，高出了 50 分。如果我们同意让一些经济困难的学生根据州政府支持的特定方案入学，而不把这些学生的 SAT 分数计入平均，当然没人会抱怨，对不对？纽约大学(New York University)就是这么做的。

数字彼此之间是否相符？

麦纳玛拉卖了通用汽车，因为通用汽车没有把他的数字和别的数字做比较。在整个改装厢型车行业一个月只生产 17 000 辆车，而且只有比 1% 多一点的车外销的情况下，竟没有人问，为什么会有车商在一个月內，可以卖 17 000 辆厢型车来外销？讲到通用汽车，下面还有一个例子，是关于数字彼此之间有点对不上的问题。

例4 我们赢了！

通用汽车的凯迪拉克车曾连续 57 年列名美国豪华车的销售第一。1998 年，福特汽车的林肯车似乎一路领先，但到最后一刻却输了。《纽约时报》说：“凯迪拉克发表了好得几乎令人难以相信的十二月销售结果之后，便以超前 222 辆车的姿态后来居上，赢得胜利。”最后的计数是凯迪拉克 187 343 辆，林肯 187 121 辆。后来



通用汽车报告，凯迪拉克的销售量在一月份跌了 38%。12 月和 1 月的销售量怎么会有这样大的差别？会不会一月份的销售量有一部分被算进前一年里去了，使得凯迪拉克刚刚好以 222 辆车险胜？事实确实如此。到 5 月的时候，通用汽车承认，它在 12 月卖的凯迪拉克，比他们所声明的少 4 773 辆。

在通用汽车的例子里，因为数字和我们的期望有差距，所以我们怀疑事情不大对。下面的例子中我们确知有问题，因为数字不对。例 5 是一篇文章的一部分，内容在批评史隆－凯特林中心 (Sloan-Kettering Institute) 的一位癌症研究员，他被指控犯了科学上的弥天大罪：伪造数据。

例 5 假数据

“有一件事他倒是完成了，就是关于明尼苏达老鼠实验的总结论文……这篇论文由史隆－凯特林中心通过，而且《实验医学期刊》(*Journal of Experimental Medicine*)也接受了。论文中有一个统计表，里面有很明显的错误，这样的错，聪明的小学生都能看得出来。表里面有 6 组动物，每组各 20 只，并包含每组成功的百分比。显然 20 的任何百分比都应该是 5 的倍数。森莫林(Summerlin)所记录的百分比却是 53、58、63、46、48 以及 67。”

数字可信吗？

正如通用汽车例子所说明的，你常常只因为数据看起来实在不可信，就查出可疑数字。有时候你可以用诸如《美国统计精粹》(*The*



Statistical Abstract of the United States)等可靠信息来源的数据,来对比检查不大可信的数字。有时候,就如下个例子说明的,你可以做些计算来证明某个数字不可能是正确的。

例6 高产瓜田

极富声望的《科学》期刊某期有一篇文章在谈论侵害植物的昆虫,内文提到加州有一块田,每英亩生产750 000颗瓜。

有读者回应:“我从小在农场长大,我知道一英亩等于43 560平方英尺,所以这块神奇的瓜田每平方英尺约可生产17颗瓜。若这些瓜是指哈密瓜,一颗就要占地接近一平方英尺,我猜它们一定是一颗叠着一颗长,总共有17层。”该读者做的计算如下:

$$\text{每平方英尺的瓜数} = \frac{\text{每英亩的瓜数}}{\text{每英亩的平方英尺数}} = \frac{750\,000}{43\,560} = 17.2$$

编辑被问得不大好意思,回答说:正确的数字应该是每英亩大约生产11 000颗瓜。

数字是否好得不像真的?

例5里面,因为数字出现矛盾状况,才使人怀疑数据是假的。而过分精确或太有规律,也一样叫人起疑,就像学生实验报告里的数据和理论结果一模一样的状况。实验助教知道,仪器的准确性和学生的实验技巧,都没有好到可以得出这么完美的结果,所以助教怀疑结果是学生编出来的。

底下是《科学》期刊中一篇文章谈到的,在医学实验中造假的例子。

在这个例子里,令人可疑的规律性(两篇论文中的数据相同)加上



不一致性(两篇论文中的动物数目不同),足以让一个细心的读者怀疑资料有假。

例 7 又见假数据

“……莱思克(Lasker)受邀写推荐函。但是在他同时读了两篇史勒次基(Slutsky)的论文后,他怀疑:两篇论文所用的控制组动物是同一批,而两篇当中却都没提到这事实。这两篇论文当中的数据完全一样,但是……两篇所引用的动物数目却不相同。这即使不是做假,也起码是非常草率的做法。当史勒次基被问到这项统计上的瑕疵之后,他几乎立刻辞职并离开了圣达戈。”

算术对不对?

错误的结论或令人无法理解的结论,常常只不过是粗心大意所造成的结果。其中,比率以及百分比尤其容易出错。

例 8 这是什么百分比

以下是澳大利亚的一些例子。《堪培拉时报》(Canberra Times)报道:“超过60岁而独居的人当中34%是女性,而只有15%是男性。”加起来只有独居人口的49%。我猜另外那51%既非女性也不是男性。

连有些聪明人也搞不清百分比。一份给女性大学教师的新闻信中间:“女性被指派为某一专业等级的机会,比男性少550%(5倍半),这样合理吗?”不管什么东西,100%就是全部了。如果拿走100%,就什么都不剩了。“少550%”是什么意思,我一点也摸不着头脑。



似乎一旦离开学校之后,很少人会再做算术。而会做算术的人就比较不会被没意义的数字给骗了。稍微思考一下,加上一个计算机,就万事搞定了。

例9 夏天多小偷

一个住宅保全系统的广告上说:“你去度假的时候,小偷就开始工作了。根据 FBI 的统计资料,有 26% 的住宅窃案发生在阵亡将士纪念日和劳动节之间。”

这样讲应该是想说服我们,小偷在暑假期间特别活跃。可是看看你的日历。阵亡将士纪念日和劳动节之间隔了 14 周。在一年 52 周当中,14 周占的百分比是:

$$\frac{14}{52} = 0.269$$

所以广告等于在说:一年当中 26% 的窃案,发生在 27% 的时间当中。这一点儿也不稀奇。

例10 老年大军来了

1976 年出版的《科学》中有位作者提出:“在美国,65 岁以上人口现在共有 1 000 万,到公元 2000 年时会达到 3 000 万,而且会占美国人口的 25%,是前所未有的高比率。”警钟响起了:老年人会在四分之一世纪里变成三倍,会构成全体美国人口的四分之一。

我们来检查一下算术。3 000 万是 1.2 亿的 25%,因为:

只不过算错一点罢了

1994 年的时候,有个由祖母级女士组成的投资俱乐部写了一本畅销书:《胡须镇女士之简易投资指南:我们如何打败股市——你要怎样做到》(*Beardstown Ladies' Common-Sense Investment Guide: How We Beat the Stock Market - and How You Can, Too*)。在书的封面上以及她们多次在电视上露面时,这些来自乡下的作者都声明她们的年度获利率是 23.4%,打败大盘以及大部分专业操盘者。四年之后一位怀疑者发现,俱乐部的会计打资料时输入错误。胡须镇女士的实际获利是 9.1%,比同期大盘的 14.9% 差很多。我们都会犯错,不过大部分的错不会像这个错这样,还能赚进大把钞票。



$$\frac{30}{120} = 0.25$$

所以 2000 年的人口总数必须是 1.2 亿，作者说的数字才说得通。美国人口在 1975 年已经是 2.16 亿了。这计算一定有什么地方不对。

既已警觉有错，我们来查一下《美国统计精粹》，看看事实如何。1975 年的时候，在美国，65 岁以上人口共有 2 240 万，而不是 1 000 万。占总人口的比率超过 10%。对 2000 年预估的 3 000 万老年人，不过是占该年预估总人口的 12% 而已。从 2000 年的立场来看，我们知道 65 岁以上人口占全美人口的 13%。人的寿命愈来愈长，所以老年人的数目会持续增加。不过是在 25 年之间从 10% 增长到 13%，比《科学》中那位作者说的，可要慢得多了。

计算某个量增加了多少百分比，或者减少了多少百分比时，很多人都会算错。一个数量改变的百分比是这样算的：

$$\text{改变的百分比} = \frac{\text{改变的量}}{\text{起始的量}} \times 100$$

例 11 股市起伏

1999 年的时候，纳斯达克指数从 2 192.69 点上升到 4 069.31 点。它增加的百分比是多少？

$$\begin{aligned} \text{改变的百分比} &= \frac{\text{改变的量}}{\text{起始的量}} \times 100 \\ &= \frac{4\,069.31 - 2\,192.69}{2\,192.69} \times 100 \\ &= \frac{1\,876.62}{2\,192.69} \times 100 = 0.856 \times 100 = 85.6\% \end{aligned}$$

表现得很突出。当然，股市有起就有落。在 2000 年 4 月 14 日往回数的一周



内, 纳斯达克指数从 4446.17 点跌到 3321.29 点, 一共跌掉的百分比是:

$$\begin{aligned}\text{改变的百分比} &= \frac{\text{改变的量}}{\text{起始的量}} \times 100 \\ &= -25.3\%\end{aligned}$$

记得分数的分母永远是要用起始的量, 而不是用较小的那个数。

一个量可以无限地增长, 增加 100% 只不过代表它变成原来的两倍。但是全世界没有什么量可以减少超过 100%, 减少 100% 就已经全部没有了。

背后有什么该注意的吗?

很多人对于各式各样的议题有强烈立场, 强烈到希望看到的数字可以支持他们的立场, 通常只要他们很小心地选择数字来报道, 或者努力想办法把数字挤压成想要的形状, 就可以找到支持他们立场的数据。以下是两个例子。

例 12 女性的心脏病

公路边一个大广告牌上简短地写着: “死于心脏病的人当中, 有一半是女性。”这个真实叙述的背后, 到底藏着什么目的? 也许立广告牌的人只是要女性知道, 她们也应注意心脏病的风险。(抽样调查显示, 许多女性低估了心脏病的风险。)

从另一方面看, 也许广告主要想反击对男性心脏病的过度强调 (有些人这样认为)。如果是这样的话, 我们也许应该指出, 虽然死于心脏病的人有一半是女性, 然而她们的平均年龄却比男性大很多。大致来说, 全美每年有 36 000 位 65 岁以下的女性及 85 000 位 65 岁以下的男性死于心脏病。美国心脏病协会 (American Heart Association) 说: “女性死于冠状动脉心脏病的风险, 差不多和小她 10 岁的男性一样。”



例 13 收入差距

在 20 世纪 80 年代和 20 世纪 90 年代美国经济起飞的时候,最高收入群和最低收入群的差距加大了。1980 年的时候,最低收入的五分之一住户,只赚到全美总收入的 4.3%,而前五分之一高收入户得到 43.7%。到 1998 年,后五分之一的低收入户,收入下降到整体的 3.6%,而前五分之一则上升到 49.2%。也就是说,前五分之一高收入户的所得,几乎是最低收入的那五分之一的 14 倍。

我们有没有办法处理一下这些数字,把差距缩小呢?《福布斯》(Forbes, 主要供有钱人阅读的杂志)里面有篇文章就这样做了。首先,平均来说,富住户的一户人数比穷住户的多,所以我们应该改成算每个人的收入。有钱人缴比较多的税,所以就考虑税后收入。穷人有食物代券及其他补助,也应该算进去。最后,收入高的人工作小时数比收入低的人多,所以应该根据工作小时数调整。这样重算之后,前五分之一高所得的收入,变成只是后五分之一低所得的 3 倍。当然啦,工作小时数较低,有可能是因为生病、身体残障、照顾孩童和年迈父母等各种原因。如果《福布斯》的背后动机是要说收入差距不重要,我们可不一定同意。

此外,还有一些地方可以做调整。普查局的收入数字当中,并不包括资本收益,比如说股票上涨卖出所赚的钱。资本收益绝大多数都进了有钱人的口袋,所以把这项加进去会使差距更大。但《福布斯》可没这样做。普查局说,如果把任何可以想像得出的,可以叫做收入的东西都调整进去的话,1998 年最低所得的五分之一住户的收入占总收入的 4.7%,而最高的五分之一占 45.8%。

网络寻奇

《美国统计精粹》是重要的资料汇编。你可以上网在美国普查局的网站上找到它。1999 年版是在 www.census.gov/statab/www/。要找你想要的资料,可以先查索引。

达特茅斯学院的 CHANCE 网站有很多有趣的东西(至少假如你对统计有兴趣的话)。其中的 *Chance News* 部分的网址为:
www.dartmouth.edu/~chance/chance-news/news.html, 上面有每个月的新闻信息,会追踪新闻界的统计资料,包括一些令人起疑的数字。



本章重点摘要

统计的目的是通过数字来洞察内情。仔细观察数字的人最有机会有所斩获。特别留意自发性回应样本及交叉问题。问清楚一个数字到底量度的是什么，并且判断一下它是不是有效量度。找一找数字的相关内容，看看是不是少了**重要信息**。看看有没有不一致的情况，也就是数字之间不如预期那样“相符”，还要检查**算术**对不对。把看来不可信的数字，即太大或太小到令人惊讶的数字，和你已知正确的数字对比一下。如果数字太有规律，或者太符合作者的愿望，也要小心。如果你怀疑某些数字被提出来，是为了支持某种**隐藏的目的**，更要仔细推敲一下。如果你养成仔细检查数字的习惯，你的朋友很快就会觉得你很有头脑，而且他们还可能是对的呢。



第9章 习题

9.1 酒后驾驶、报纸上一篇探讨酒后驾驶的文章引用了罗得岛的车祸死亡资料：“死亡数有 42% 发生在周五、周六及周日，显然是因为周末饮酒的人比较多。”周五、周六加上周日占一星期的百分之几？那你会不会惊讶有 42% 的死亡事故发生在那 3 天？

9.2 止痛药广告、止痛药泰诺(Tylenol)的一则广告用了这样的标题：“为什么推荐泰诺的医师，比推荐所有知名品牌的阿司匹林加起来还要多。”拜耳阿司匹林的制造者，在以“泰诺制造者，你真可耻！”为标题的回应中，指控泰诺只提供部分事实，误导民众。你来当侦探，告诉我们为什么泰诺的声明即便是事实仍然误导？

9.3 止痛药广告、阿纳辛(Anacin)长久以来的广告都说：“在最受医师推荐的成分中，它包含得比较多。”另一种在药房不用处方就可买到的止痛药声称：“医师最常指名百服宁，而不是其他的‘知名品牌’。”两种广告词在字面上都是正确的：美国联邦贸易委员会(Federal Trade Commission)却认为二者都误导。请说明原因。(提示：想一想，阿纳辛和百服宁当中的主要止痛成分是什么？)

9.4 郊区的鹿群 西彻斯特县是纽约市北边的郊区，面积共 438 平方英里。某田园杂志宣称，该郡土地上共有 800 000 只鹿。做个算术来证明这个数字不可信。

9.5 越战老兵的自杀情况 越战的恐怖经验，是不是驱使许多参与该役的美国退伍军人走上自杀之路？有一个被广泛引用的数字是，在战争结束后的 20 年间，共有 150 000 名越战退伍军人自杀。说明为什么这个数字不可信？以下一些事实可以帮你判断：每年约有 25 000 名美国男性自杀；越战期间约有 300 万男性在东南亚服役；美国成年男性约有 9300 万。

9.6 大海中的垃圾？一篇谈论游轮丢垃圾污染海洋的报道中说：
一艘中型游轮(约 1 000 位乘客)的海上七日游，可能



累积了 222 000 个咖啡杯, 72 000 个汽水罐, 40 000 个啤酒瓶或罐, 以及 11 000 个酒瓶。

这些数字可信吗? 做些算术来支持你的结论。假设船员人数和乘客一样多。照这个数字算来, 每个人每天必须要喝多少杯咖啡?

9.7 奇怪的数字。以下是某科学期刊中一篇书评中的一段:

……总共 20 项研究当中, 有 57% 报告了有统计显著意义的结果, 其中有 42% 同意某一项结论, 而另外的 15% 同意另一项结论, 通常是相反的结论。

这段话里的数字合不合理? 你能不能算出在那 20 项研究中, 有几项是同意“某一项结论”, 多少项同意另一项结论, 而多少项没有具统计显著意义的结论?

9.8 机场的延误状况。中西部一家报纸报导主要机场航班延误情况, 说道:

根据甘尼特新闻服务公司对于过去五个月之间美国航空公司表现的研究结果得知, 芝加哥的欧黑尔机场原本排定了 114 370 个航程, 但有将近 10%, 共 1136 个航班被取消了。

检查一下这个报纸的算术。欧黑尔机场的预定航班, 有多少百分比被取消了?

9.9 挨打的女人? 有人写信给《纽约时报》的编辑, 对一篇时报社论提出异议, 社论中说: “每 15 秒钟, 就有一位美国女性被她的先生或者男朋友打。”写信的人说: “以这种比率来算, 每年会有 2 100 万个女性被先生或男朋友打。这当然不是事实。”他引用了全国罪案受害者调查的估计, 丈夫施暴件数有 56 000 件, 男友或前男友施暴件数有 198 000 件。调查指出对女性的攻击事件总共有 220 万件, 大部分的加害人是陌生人, 或者这个人受害者虽然认识, 却非丈夫、男友或前夫、前男友。

- (a) 先做一下算术。每 15 秒一件等于每分钟 4 件, 以这种比率来算, 一个钟头里有几件打人事件? 一天呢? 一年呢? 写信的人算术对不对?
- (b) 写信者指控时报夸大了女性受侵害的家庭暴力事件件数, 立论是否正确?



9.10 我们能读，但是能算吗？美国普查局有一次让一个3400人的随机样本考了简单的英文识字测验，《纽约时报》登了部分考题，用的标题是“113%的美国成人没有通过这项测验”。为什么标题中的百分比显然错了？

9.11 股票下跌。1998年8月4日那一天，这环指数从开盘时的8780.94，下跌了299.43点。有些报纸的标题说这是“有史以来点数降幅的第三大”。那天道琼斯跌了多少百分比？因为指数在前一年上升非常多，所以跌的百分比连前20名都排不上。比率比计数更容易看得清楚，这是又一个例子。

9.12 贫穷。生活低于官方贫穷标准的美国人数，在1979—1998年的20年之间，从26072000增加到34476000。上升了多少百分比？你不应该根据这些数字就做出结论，认为贫穷的情况愈来愈普遍。为什么？

9.13 我们不会把你的行李给丢了。大陆航空公司有回做广告说，他们“在过去六个月当中，把找不到行李的情况减少了100%”。你相信他们说的吗？

9.14 用漱口水漱口，有一种新的漱口水声称可以“把牙菌斑减少300%”。仔细解释一下，为什么不可能把任何东西减少100%。

9.15 校园中的希腊人。一份校园报纸的问答专栏中有人问，学生当中有多少百分比是“希腊人”（指兄弟会或姊妹会的成员）。专栏的回答是：“以第一学期的情形来说，女生约有13%，男生在15%—18%之间，因此‘希腊人’占全体大学部学生的百分比，大约是在28%—31%之间。”讨论一下这份大学报的算术算得如何。

9.16 不敢开车？美国有所大学每个月寄一份有关健康的新闻信给它的员工。最近一期里有个叫做“机会有多大”的专栏里说：

你今年会死于车祸的概率是：1/75

美国人口大约是2.75亿，其中每年有大约40000人因车祸而死亡。随便选个人，他今年会死于车祸的概率是多少？



9.17 有多少英里的公路?《有机种植》(*Organic Gardening*)杂志有一次说:“美国州际公路系统共长 390 万英里,而且损耗的速度比修复速度快 50%。路面状况持续恶化,使驾驶人每年增加了 70 亿的燃料支出。”美国东西岸的距离约为 3 000 英里。一共要有多少条横跨美国的公路,才能得到总长 390 万英里?你觉得 390 万英里这个数字正确吗?

9.18 在花园里 《有机种植》杂志在描述怎样可以改变你花园里的土质时说:“面积 100 平方英尺厚度 6 英寸的土层,重量约 45 000 磅,只要加进 230 磅的堆肥,会立刻让你有 5% 的有机物质。”

(a) 230 是 45 000 的百分之几?

(b) 水的重量约为每立方英尺 62 磅。花园里面积 100 平方英尺,厚度 6 吋的一土层,体积总共是 50 立方英尺。50 立方英尺的水应该有多重?你觉得 50 立方英尺的土总重量 45 000 磅可信吗?

(c) 从(b)中可以看出,45 000 磅这个数字不正确。事实上,土的重量约为每立方英尺 75 磅。如果我们用正确重量来算,“5% 有机物质”的结论大致正确吗?

9.19 没有合适的男人?某新闻报道引用了一位社会学家的话,说在美国对应于每 233 位未婚的 40 多岁女性,40 多岁的未婚男性只有 100 位。从这些数字可以想像出该年龄层妇女尴尬的社交处境。这些数字可信吗?(《美国统计精粹》里有一个表的标题是“根据年龄和性别分类的婚姻状况”,里面的数字可以用来参考。)

9.20 好得不像真的?已过世的英国心理学家伯特(Cyril Burt)以研究分隔两地成长的同卵双胞胎的智商而著称。伯特研究中,分隔两地的同卵双胞胎 IQ 的高相关系数(correlation),显示 IQ 主要是受遗传影响(“相关系数”度量两个变数之间关系有多紧密。我们会在第 14 章中讲到相关系数)。伯特对他的研究结果写了多次报告,纳入研究的双胞胎对数也随着时间增加。以下是他发表结果时所报告的相

发表日期	分隔两地成长的双胞胎	一起成长的双胞胎
1955 年	0.771(21 对)	0.944(83 对)
1966 年	0.771(53 对)	0.944(94 对)



关系数。

有什么令人起疑之处吗?

9.21 从哪开始有差别 当要比较随着时间改变的数字的时候,你可以通过选择起始点,来使比较的结果倾向不同方向。比如说芝加哥小熊队先输 5 场球,再赢 4 场,再输 1 场。你可以诚实地说小熊队在过去 10 场球中输了 6 场(听来不大妙),或者说他们在过去 5 场球中赢了 4 场(听来很棒)。

美国家庭的收入中位数(median)(根据持续购买力,以美元为单位)1989 年为 44 284 美元,1993 年的 41 051 美元以及 1997 年的 44 568 美元。在 1989 和 1997 年之间,家庭收入增加多少百分比?1993—1997 年之间呢?你会看到,你可以通过选择不同的起始点,来让收入的变化看起来很好或是不好。

9.22 是否高薪也有差别 上一题习题指出,家庭收入中位数在 1989—1997 年间几乎没有变化。收入最高的前 5% 住户,在 1989 年至少赚了 120 607 美元,1997 年至少赚了 128 521 美元。(数字根据持续购买力,以美元为单位)。在 1989 年—1997 年之间,高收入者的收入增加了多少百分比?

9.23 行船安全 《美国统计精粹》当中,有关休闲性质的行船意外事故的资料显示,死亡人数由 1980 年的 1 360 降到 1990 年的 865,再降到 1997 年的 819。然而受伤人数却从 1980 年的 2 650 增加到 1990 年的 3 822,然后到 1997 年的 4 555。在这些政府资料中,为什么相对于受伤人数,死亡人数这么少?哪个计数(死亡人数或受伤人数)可能比较准?为什么有可能受伤人数上升,死亡人数却下降?

9.24 请假 迪尔伯特卡通中的尖头发老板有次注意到,员工请的病假有 40% 是在周五或周一。这是不是员工想要休长周末的证据?

9.25 为下列每一项找个例子。从统计观点详细解释你的例子的缺失。

漏列重要信息



数字之间不相符

不可信的数字

错误的算术

可以找到资料的地方之一，是《网络寻奇》里提到的网上
Chance News。

第一部分 复习

对任何统计研究应该要问的第一个，也是最重要的问题，是“数据从哪儿来的？”第 1 章谈了这个问题。观测所得和实验数据的差别，是答案中的关键部分。好的统计始于产生数据的好设计。第 2、3、4 章讨论抽样，这是从总体选一部分来代替个体的艺术。图 1.1 综合了简单随机样本的主要概念。第 5 及第 6 章考虑实验设计的统计面，实验是指执行某些处理以便了解所产生的反应的研究。随机化比较实验是其中的重要概念，图 1.2 描绘了最简单设计的概要。

随机样本和随机化比较实验，可能是 20 世纪最重要的统计发明。两者都经过很长的时间才被接受，而且你仍将会看到很多自发性回应样本以及没有控制组的实验。随机样本和随机化实验都刻意用机遇来消除偏差，并产生有规律形态的结果。这个规律形态让我们有办法可以算出误差界限，做出置信叙述，并且对根据样本或实验做出的结论，评估其统计显著性。

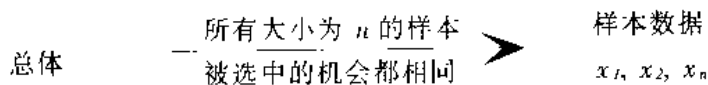


图 1.1 简单随机样本的概念

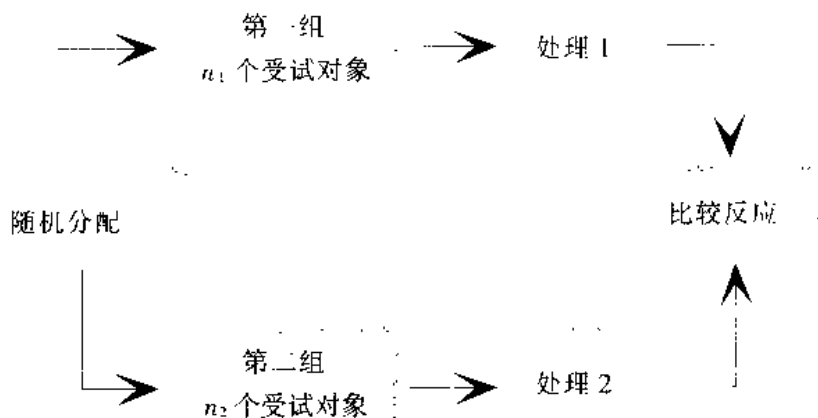


图 1.2 随机化比较实验的概念

当我们搜集关于人的数据时，伦理议题可能会很重要。第 7 章讨论这些议题，并介绍了三项原则，适用于任何以人作为受试对象的研究。产生数据的最后一步，是度量我们感兴趣的特征，以产生数字供我们研究用。量度是第 8 章的主题。“数据从哪儿来的？”是我们对一项研究该问的第一个问题，而“数字看来合理吗？”是第二个。第 9 章鼓励大家养成一个好习惯，那会很有收获，就是在接受数字似乎是在说什么之前，要先用怀疑的态度检视一番。



第一部分 重点摘要

以下是你读完 1—9 章后，应该要有的最重要的技能。

A. 数据

1. 能辨认出一项统计研究当中的个体和变量。
2. 能分辨出观测研究和实验。
3. 知道什么是抽样调查、普查及实验。

B. 抽样

1. 能指出抽样的总体。
2. 看得出自发性回应样本和其他坏抽样方法所产生的偏差。
3. 会用表 A 的随机数字表，从一个总体中选出简单随机样本。
4. 能解释抽样调查在做结论时，怎样对付偏差及变异性。用日常用语来说明，一项抽样调查结果的误差界限以及“95% 的信心”是什么意思。
5. 能用速算法找出 95% 置信水平的近似误差界限。
6. 了解抽样误差和非抽样误差的差别。看得出涵盖不全和不回应抽样调查中的确是误差来源。了解问题的措辞对回答的影响。
7. 在已决定如何分层之后，会用随机数字从一个总体选出分层随机样本。

C. 实验

1. 可以在一项实验中分辨出什么是解释变量、处理、反应变量以及受试对象。
2. 不论在观测研究还是实验当中，能看得出由于解释变量和潜在变量的交叉所产生的偏差。
3. 会用像图 I.2 那样的图来描绘完全随机化实验的设计。在图里面要把分组的大小、所有处理及反应变量都表示出来。



4. 会用表 A 的随机数字，来执行完全随机化实验当中，把受试对象随机指派到各组这一部分。
5. 会在合适的时候用配对或区集设计。
6. 对安慰剂效应有认知。知道什么时候应该用双盲方法。了解实验的弱点，尤其是结果能不能推广。
7. 能解释为什么随机化比较实验可以对因果关系提供充分证据。
8. 能说明何谓统计显著性。

D. 其他主题

1. 会说明数据伦理的三项首要原则。并讨论在特定情况下，这些原则要怎么用。
2. 会说明在特定情况下，怎样可以利用度量来把变量定义清楚。
3. 能够评量一个变量做为某一特征的量度，是否具备有效性，包括预测有效性。
4. 能解释在度量时，如何可以降低偏差及加强可靠程度。
5. 认得出不彼此之间相符的数字、不可信的数字、好到令人起疑的数字以及错误的算术。
6. 会正确计算增加的百分比及减少的百分比。



第一部分 复习习题

复习习题都简短易答，目的是要加强你在书里学到的基本概念和技巧。

I.1 了解这些用词。有位不懂统计的朋友在她的一门心理学课程的阅读作业当中，遇到了一些统计名词，用一两句简单的句子说明下面的每一项名词。

- (a) 简单随机样本。
- (b) 95% 置信水平。
- (c) 非抽样误差。
- (d) 知情且同意。

I.2 了解这些用词。有位不懂统计的朋友在她的生物课里碰到一些统计用语。用一两句简单的句子来说明下面各项名词。

- (a) 观测研究。
- (b) 安慰剂效应。
- (c) 有统计显著意义。
- (d) 试验审查委员会。

I.3 有偏的样本。你看到一位女同学站在活动中心前面，不时拦下其他学生来问问题。她说正在为一份课堂作业搜集同学的意见。说明为什么这种抽样方法几乎一定是有偏的。

I.4 选取 SRS。一所规模很大的大学中，有一位学生想要研究，当学生打电话到各系询问事情时，会得到怎样的回应？她要从以下的学系清单中，选出 6 个系的 SRS 来研究。请从表 A 的列 116 开始，帮她选这个 SRS。

农艺	教育	护理
艺术设计	电机	药学
听力学	英文	哲学
生物化学	外语	物理



生物	历史	政治
化学	园艺	心理
传播	工业工程	社会
计算机科学	管理	统计
消费科学	数学	动物解剖学

1.5 选取 SRS。一所大学的教职员申诉系统,明白规定要从 30 位成员的申诉委员会中随机选取一个 5 人听证小组,从表 A 的列 107 开始,自下列申诉委员名单中,选取一个 5 人的 SRS。

埃布尔	福利	洛	桑多葛洛奇	沃尔瑟
阿德勒	冈萨雷斯	马丁内斯	施拉	温斯坦
考尔德伦	海	米勒	史岱史斯	韦纳
迪维托	伊斯梅尔	莫克	斯蒂尔	吴
艾辛格	柯比	伯鲁奇	塔格	扬
尤班克	利维	罗森伯格	特欧	张

1.6 抽样调查的误差 举例说明抽样调查中的非抽样误差来源。再举一个抽样误差来源的例子。

1.7 抽样调查的误差 一项抽样调查随机选择电话来从事访问,这种抽样方法会漏掉所有没装电话的人。这是非抽样误差还是抽样误差的来源?调查结果中宣布的误差界限,有没有把这种误差来源考虑进去?

1.8 抽样调查的误差 有一所大学从大学部学生注册名单当中,随机抽取了 100 个学生的 SRS 来访问,听取他们对大学生活的意见。若他们同时抽取了两个 100 人的 SRS,从这两个样本得出的结果,恐怕会有些不一样。这种变异是抽样误差或是非抽样误差的来源?调查结果中宣布的误差界限,有没有把这种误差来源考虑进去?

1.9 抽样调查的误差 习题 1.7 和 1.8 各提到抽样调查中的一个误差来源。如果把抽样大小加倍,但是其他抽样步骤不变,可以减少上述的误差吗?要说明理由。



I.10 抽样调查的误差。盖洛普调查发现,有 68% 的美国成人,赞成在公立小学中同时教神创论和进化论。盖洛普的新闻稿说:

对于根据这样大小的抽样所得的结果,在 95% 的置信水平之下,我们可以说,由于抽样及其他随机效应所产生的误差,会在正负 3 个百分点范围内。

举一个例子指出这项民调结果的误差来源,而且这个误差来源是有没有包括在误差界限之中的。

I.11 找出误差界限。一项调查访问了 586 位在上周曾使用互联网的成人,并问了下列问题:“互联网是让你的生活好得多、好一点、差一点、差很多,还是对你的生活没影响。”586 位受访者中总共有 152 位说“好得多”。

(a) 这项抽样调查的总体是什么?

(b) 用速算法算出误差界限。然后用一个完整的置信叙述对总体做结论。

I.12 找出误差界限。如果低收入户想送小孩上私立学校或教会学校,政府应该提供补助吗?一项对 1 006 位成人做的调查,结果是有 362 位赞成。

(a) 这项抽样调查的总体是什么?

(b) 用速算法算出误差界限。对总体的意见提出一个置信叙述。

I.13 什么样本?一个聚会当中有 30 个满 21 岁的学生以及 20 个不满 21 岁的学生。你从满 21 岁的当中随机抽出 3 个,另外从不满 21 岁的当中随机抽出 2 个,然后问他们对酒精的看法。你给了参加派对的每位学生同样被访问到的机会,请问这个机会是多少?你的抽样为什么不是 SRS?这种抽样叫什么?

I.14 设计实验。一所大学的统计系希望多吸引些学生来主修统计。该系准备了两种宣传小册,小册 A 强调统计可以提供的激情,小册 B 强调统计学家可以赚多少钱。到底哪一样比较能吸引大一学生呢?现在你有一项问卷,可以度量学生主修统计的意愿,而且还有 50 个大一新生共同参加这项研究。大略描述一下,怎样设计实验来分辨哪个小册子的效果较好。



1.15 设计实验 盖瑞认为,要让女孩愿意跟你约会,方法是先向对方谈自己。葛雷却认为,让对方谈她自己,效果反倒更好。你招募到20位两种谈话方式都愿意试的男生,谈话隔天叫他们打电话给对方要求约会。大略描绘一项实验设计,可以决定哪个方法比较会成功。

习题 1.16—1.19 根据的是《美国医学会期刊》中的一篇文章,该篇文章谈的是流感疫苗是否有效。文章中报道了对于一种叫做 trivalent LA IV 的鼻喷剂疫苗有效性的研究结果。以下是文章的部分摘要:

设计 随机化,双盲,有安慰剂控制组,试验进行时间为1997年9月到1998年3月。

参加者 一共有4561位身体健康、有工作的成人参加,年龄在18—64岁之间,这些参加者是经由健康保险系统、工作场所,以及从一般大众中招募的。事件参加者于1997年秋以2:1的比例,经由随机选择,接受了鼻喷剂的 trivalent LAIV 疫苗($n=3041$)或安慰剂($n=1520$)。

结果 疫苗导致……工作天损失减少了(严重发烧减少了17.9%;发热性上呼吸道疾病减少了28.4%),也减少了看病天数(严重发烧减少了24.8%;发热性上呼吸道疾病减少了40.9%)。

1.16 了解这些用词 在该研究的设计描述中,“随机化”、“双盲”和“有安慰剂控制组”是什么意思,各用一句话解释。

1.17 实验基础知识 指出这项研究中的受试对象、解释变量及数个反应变量。

1.18 设计实验 用图来描绘这项医学研究的实验设计。

1.19 道德 数据伦理的三项首要原则是什么?简短说明在这项流感疫苗研究中应该怎么做,才能符合这三项原则。

1.20 度量 琼妮想要量量看,大学男生在政治上的“左”倾程度。她决定量他们的发长——头发愈长愈“左”倾。

(a) 这个方法看来可靠吗?为什么?

(b) 这个量度看来无效,为什么?



(c) 然而, 用发长来度量政治倾向, 可能有一些预测有效性。解释一下怎么会这样。

1.21 可靠程度 你在努力做化学实验, 指定的作业是要度量一种溶液的传导性。说你的量度很可靠是什么意思? 怎样可以改进你实验结果的可靠程度?

1.22 是观测还是实验? 美国“护士健康研究”(Nurses Health Study) 从 1976 年以来, 每两年就会向超过 100 000 位有执照的女护士样本问问题。该研究从 1980 年开始问饮食习惯, 包括饮不饮酒。研究者的结论是“少量或适度饮酒的人, 死亡的风险具统计显著性的低于不喝酒或大量喝酒的人”。

(a) 护士健康研究是观测研究还是实验? 为什么?

(b) 在统计报告中, “有统计显著性”是什么意思?

(c) 想想看, 有哪些潜在变量, 可能可以解释为什么适度饮酒的人比不喝酒的人死亡率还低(该项研究有针对潜在变量做调整)。

1.23 是观测还是实验? 有一项针对健康状况和个性之间关系的研究当中, 自愿参加某项运动计划的中年大学教职员, 根据健康检查的结果, 分成健康状况较差及健康状况较佳组。然后所有受试者参加了一项个性测验。健康状况组在“自信”一项有较高的平均成绩。

(a) 这是观测研究还是实验? 为什么?

(b) 我们不能下结论说, 健康状况较佳造成较高的自信。想想看, 这些变量之间, 以及可能和其他潜在变量之间还有什么其他关联, 是可以解释为什么健康状况较佳组的自信心较高?

1.24 百分比的增减。在 1999 年 1 月—2000 年 3 月间, 加州普通汽油的平均价格, 从每加仑 1.2 美元增加到每加仑 1.8 美元。

(a) 验证一下, 这一来价格上升了 50%。

(b) 如果汽油价格从 3 月的每加仑 1.8 美元降低了 50%, 则新价格是多少? 请注意先增加 50%, 再减少 50%, 并不会让我们回到原点。

1.25 减少的百分比 2000 年 4 月 4 日那个星期二, 纳斯达克指数开盘时为 4 283, 然后中午过后不久曾跌到 3 649。指数减少了



多少百分比?

1.26 不可信的数字?《新闻周刊》(*Newsweek*)有一次在一篇报道中说,目前单身的40岁左右的女性,被恐怖分子杀掉的概率比结婚的概率高些。你觉得可信吗?哪一类资料可以帮你检验这项声明的正确性?



第一部分 报告作业

报告作业是比较长的习题，需要搜集信息或制作数据，而且重点是要把做出的结果用一篇短文来说明。这里很多题目适合由一组学生共同来做。

作业 1. 自己设计抽样调查 选一个你同校同学目前感兴趣的议题。准备一份简短(不超过 5 个问题)的问卷，来搜集对这个议题的意见。选出一个大约 25 人的学生样本，让他们填答你的问卷，并且简短描述从问卷结果中发现了什么，并且把你在设计和执行调查时的经验，用一段文字稍做讨论。(虽然因为 25 个学生太少，使你从统计角度没法对结果有太多信心，但这题作业的重点在抽样调查的实际执行层面。你必须先界定一个总体：如果较大的学生总体无法掌握，就用正在修这门课的同学亦可。填答问卷的人觉得你的问题清楚吗？你写问题的方式，是不是使得整理结果的时候很容易？问卷做完以后，你有没有希望自己问了不一样的问题？)

作业 2. 度量 习题 6.13 要求你执行一个配对实验的设计，来了解惯用右手的人是否右手比左手有力。你可以这样子度量手力：把一个家用体重计放在一个架子上，一端超出架子，再叫受试者用手去挤压突出架子的部分。找几个受试者来试，然后判断一下这样量手力是否可靠。写下并说明你的发现。比如说，你是否发现大家抓体重计的方式不一样，因此如果要让度量方法一致，应该要有详尽的指示。为受试者写一份这样的指示。

作业 3. 实验 在你或者你的小组，把前一题当中怎样量手力的细节决定之后，用至少 10 位受试者来执行习题 6.13 的配对实验。写一份报告，内容要描述如何随机化，列出实验数据，报告两手手力差距(右手减左手)，并且给个结论，你的小小实验是否显示，平均来说右手比较有力。

作业 4. 描述一项医学研究 访问《美国医学会期刊》(JAMA)的网站(jama.ama-assn.org)。JAMA 和《新英格兰医学期刊》的网站



(www.ncbi.nlm.nih.gov)不一样, *JAMA* 让大家可以在网上免费阅读期刊中文章的全文。从最新一期期刊, 或者过去任一期期刊中, 选一篇讨论某个研究的文章, 其研究主题是你感兴趣的。写一篇报纸报道, 把研究的设计和发现摘要写出来。(一定要包括统计学方面的说明, 比如是观测研究还是实验, 以及有没有用到随机化。新闻报道中通常忽略这些事实。)

作业 5. 国产车和进口车 在你就读的大学中, 学生和教职员比起来, 开进口车的机会比较大还是比较小? 设计并执行一项研究来找出答案, 并写一份报告说明你的设计和你的发现。你得先把国产车和进口车定义清楚, 使每一辆车在归类时不会模棱两可。然后你得找到合适的汽车样本——也许可以到学生停车区和教职员停车区去找。如果停车区域很大, 你不必一辆一辆去看, 取样本就好了。可以用系统样本(习题 4.25)。

作业 6. 数据伦理 找一篇新闻报道, 其内容和统计研究的伦理议题有关。对这项争议做一个摘要, 并写出你认为可以做出的结论。

以下是怎样着手这个作业的一个例子。在我正在写这段内容时, 就有一项争论正在进行当中, 争执不下的是, 在非洲国家研究艾滋病, 却没有把已经对富有国家提供的很花钱的疗法用在每一个受试者身上, 这到底合不合乎伦理? 到《纽约时报》网站 www.nytimes.com 的档案里, 以“艾滋和伦理”(AIDS and ethics)为关键字寻找, 会找到很多篇文章, 包括 2000 年 3 月 30 日那天一篇很好的文章。要读上面的文章, 你必须付 2.5 美元, 不然就要去图书馆查阅。

作业 7. 度量收入 一个住户的“收入”指什么? 住户收入可以用来判断, 够不够资格被纳入政府的协助低收入户计划。收入的计算方式还牵涉到政治效应。政治保守派常常声称政府资料把穷人的入数给夸大了, 因为资料里面只包括金钱收入, 没有计入食物代金券的等价金额以及房屋津贴。自由主义者则回应说, 政府应该只计入金钱收入, 这样才看得出有多少人需要帮助。

假设你是一位美国国会议员的助理, 这位议员正在考虑一项新的福利法案。写出“收入”的确切定义, 当做判断哪些住户够格接受福



利补助的依据，且必须写成一篇短文。你会不会把非金钱收入，例如食物代金券的价值以及房屋津贴也包括进去？为了让父母能够外出工作而必须花费的托儿支出，你会不会从收入中扣掉？有些值不少钱，但并不增加收入的资产，比如房子，又怎么算？

第二部分

整合数据

文字本身不构成故事。得由作者把文字组成句子，再把句子组成故事。如果文字组合得不好，故事可能让人看不大懂。数据也一样，如果要让人看清楚数字隐含的信息，一样需要经过整合。字用得太多，会让主题变模糊而不是变清楚。一大堆的数据更是叫人难以消化，因此我们常常需要一个精简的摘要，来突显出重要内容。应该如何整理、综合及呈现数据，就是本书第二部分的主题。

整理及综合大量的事实时，最容易使事实受到扭曲，其中有些是无心的，有些却是故意的。不管呈现事实用的是数字还是文字，上述情况发生的机会都差不多。我将指出，在呈现数据时会让人不小心的人上当的一些陷阱。把统计看成说谎艺术的人，看统计时注意力都放在数据的综合及呈现的那部分上。我却主张，误导的综合资料及选择性的呈现资料，早在偷食禁果之后，亚当、夏娃和上帝的谈话中就已开始了。不要怪统计。要记得那句老话：“数字不会说谎，但骗子却会。”所以要小心。

第 10 章

好的图及坏的图

有钱了

*译注：指上涨。

在 20 世纪的末期，美国股票市场是大牛市*。有多大呢？从图上面看要比用文字说明清楚得多。

先看次页的图 10.1。图中显示的是 1971—1999 年之间，每一年股价增加或减少的百分比〔用标准普尔 500 指数(Standard & Poor's 500 Index)当标准〕。在 1982 年之前股价上上下下。有时候降得很厉害，例如股票市值在 1973 年损失了 15%，1974 年又损失了 27%。但从 1982 年开始的 18 年当中，有 17 年股价都上涨，而且常常上涨很多。

从图 10.2 可以看出怎么样发财。如果你在 1970 年底投资 1 000 美



元在股票上，从图上可以看出来你的 1 000 美元在之后的每一年的年底变成多少钱。1974 年过完的时候，你的 1 000 美元掉到 854 元了，而在 1981 年底，只上升到 2 121 美元，这样差不多是每年增长 7%。在那段期间你把钱放在银行里反倒会赚得多些。然后大牛市上场了，在 1999 年底的时候，你的 1 000 美元已变成 44 875 元。

从这里可以学到的是，只要你肯长期持有，就应该把钱投入股市。可能有人学到的却是，应该在 1980 年把钱投入股市，然后在 2000 年把股票卖掉。不过我们在这里学到的统计，是针对脑袋而不是针对钱包的。而图可以很清楚呈现数据究竟说了些什么。从图 10.1 可以看出，股票的逐年表现多么不规则。图 10.2 显示出，长期下来股价大幅增长，给了有耐心的投资人很大的回馈。从 1995 年开始的 5 年当中，大幅增长的结果尤其明显。

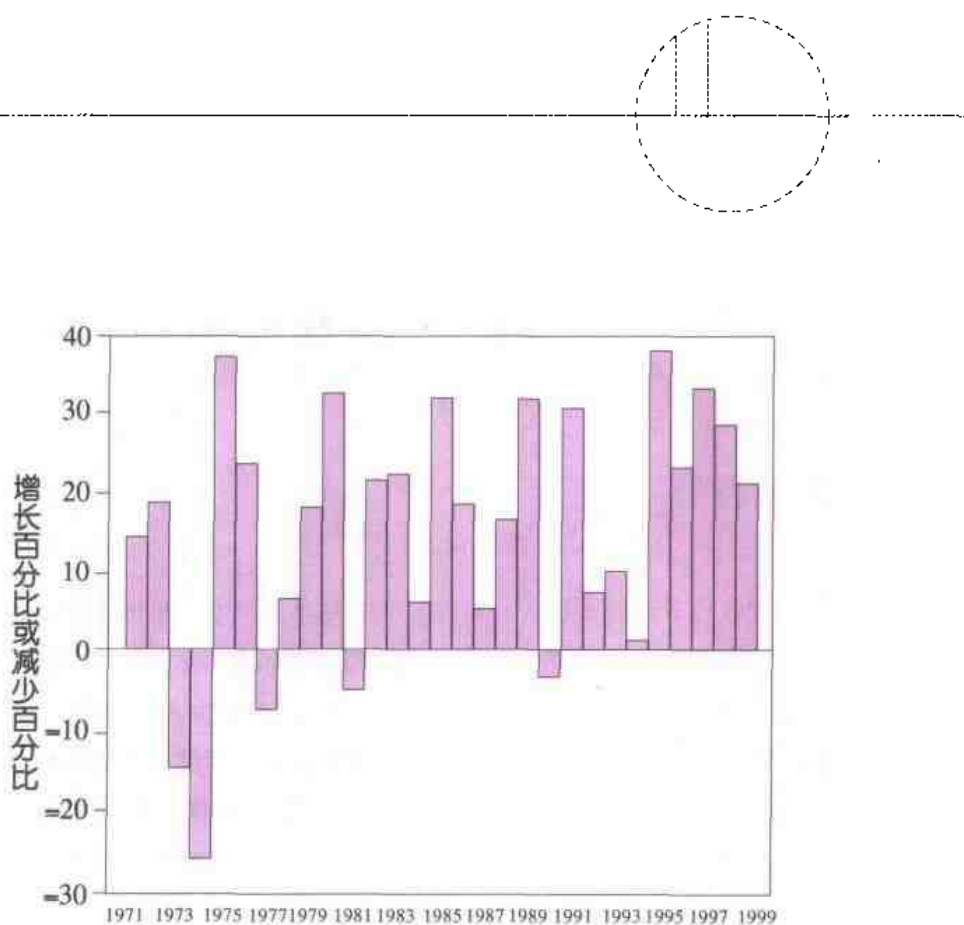


图 10.1 1971—1999 年，标准普尔 500 普通股价指数的增长百分比或减少百分比

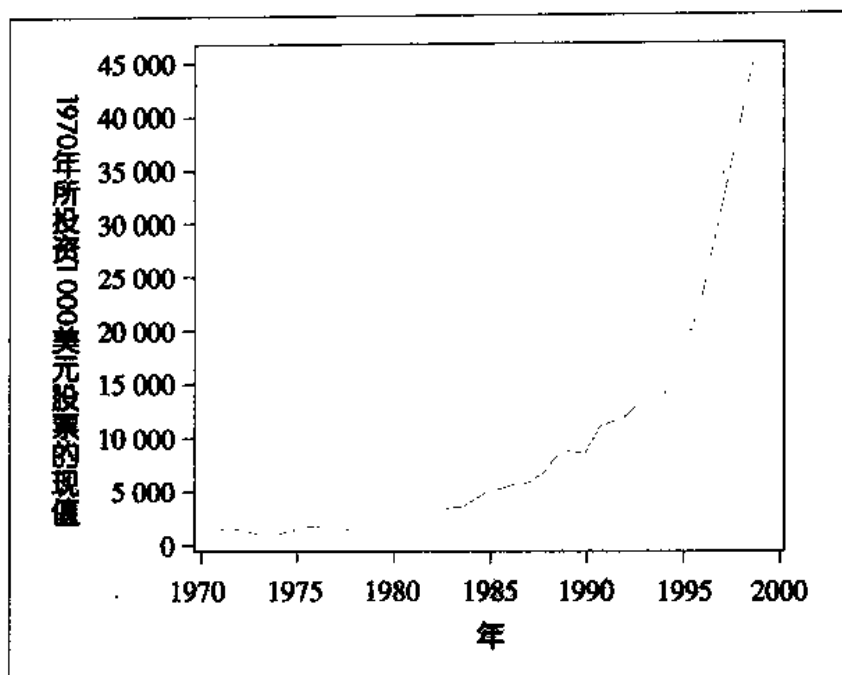


图 10.2 1970 年底对标准普尔 500 指数投资的 1 000 美元，在 1971—1999 年之间，每年年底的价值变化

听到统计这两个字时，脑袋里会出现什么画面？多半是些塞满数字的表，或者是一会儿往上一会儿往下的折线图。这样的画面没错：统计是处理数字的，而我们用表和图来呈现数字。经由随机抽样，随机化比较实验和有效量度可以生产出好的数据。现在我们要看看，数据说了些什么。

数据表

建议你看看《美国统计精粹》，它每年出一本，里面有各式各样的数值信息。私立小学和中学的数目是不是有增长？这些学校的学生中，弱势群体占多少？过去几年当中，每年共有多少人得到学士学位？而这些学位若根据学习领域来分，或者根据获得者的年龄、种族或者性别来分的话，各有多少？所有这些以及更多其他的信息都可以在《美国统计精粹》的《教育》那一节里找到。这些数据表(Data table)把资料做了摘要。我们并不想要有关每一个大学学位的信息，只想知道我们感兴趣的一些类别当中的计数。



例1 怎么样的表才清楚?

30岁左右的年轻人,教育程度如何?表10.1呈现了25—34岁的人的资料。这个表是数据表的一个好的示范。表的标示很清楚,所以资料的主题一目了然。主标题描述了资料的总主题,并且列出年份,因为这种资料会逐年改变。表里面的标题阐明了变量,并且说明了度量变量所用的单位,例如,你可以注意一下,计数以千为单位。资料来源出现在表的底部。这份普查局发表的结果,内容资料事实上是从我们的老朋友,“当前人口调查”那儿取得的。

表10.1先列出不同教育程度年轻人的计数。比率(或者百分比)通常要比计数清楚——听到说有12.1%的年轻人没有读完高中,比起听到有4 754 000个年轻人没有读完高中,前者的信息要清楚得多。表10.1中也列出百分比,表里面的这两行数字,用两种不同方式呈现了“教育程度”这个变量的分布(distribution)。每一行提供的信息,包括变量可能有什么值*,以及这个变量等于其中每一个值的比率。

*译注:指表中所列5种不同教育程度。

表10.1 25—34岁人士的教育程度,1998年

	人数(以千人为单位)	百分比(%)
高中以下	4 754	12.1
高中毕业	12 568	31.9
曾读大专	11 220	28.5
有学士学位	8 367	21.3
更高学位	2 444	6.2
总数	39 354	100.0

资料来源:普查局《1998年3月美国教育实况》。

• 变量的分布

一个变量的分布(distribution),告诉我们变量有什么可能值,以及每一个值出现的比率。



例 2 舍入误差

你有没有检查一下表 10.1 中的数字是否相符?人数的总数应该是

$$4\,754 + 12\,568 + 11\,220 + 8\,367 + 2\,444 = 39\,353 (\text{千人})$$

可是表里面的总数是 39 354。怎么回事?里面每一单项的数字,在换成以千人为单位时经过四舍五入。因为是每一项个别做四舍五入,加起来和总数不合是正常的。从此以后,这种舍入误差(roundoff error)在我们做算术的时候,都会一直跟着我们。

饼状图及柱状图

表10.1 当中的分布很简单,因为“教育程度”只分成 5 种可能值。要把这个分布用图来表示的话,可以用**饼状图**(pie chart)。图 10.3 就是年轻人教育程度的饼状图。饼状图可以显示一个整体怎样分为几个部分。要画饼状图,先得画个圆,圆代表全体,在这个例子中,就是所有 25 到 34 岁的人。圆里面的扇形就代表各部分,各扇形的圆心角和各部分的大小成比例。比如说,有 21.3% 的年轻人有学士学位但没有更高的学位。因为一个圆的内角为 360 度,所以代表“学士学位”的扇形,其圆心角就有:

$$0.213 \times 360 = 77 \text{ 度}$$

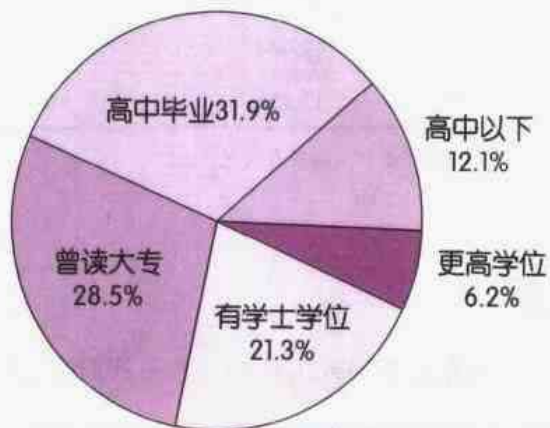


图 10.3 1998 年 25—34 岁人士教育程度分布的饼状图



饼状图的好处是让我们看到：所有的部分合起来，的确是全体。但是角度比长度难比较，所以饼状图并不是比较各部分大小的好方法。

图 10.4 是根据同样数据所画出的柱状图(bar graph)。每个柱状的高度显示出：年轻人中合于该柱状底部标示的教育程度的，占多少百分比。从柱状图可以清楚看出，高中毕业生比读过大专的人多，因为“高中毕业”的柱状比较高。而这种差异在饼状图的扇形之间不容易看出来，所以我们得在每个扇形都标示百分比。除非是用电脑绘图，否则柱状图一般来说比饼状图好画。

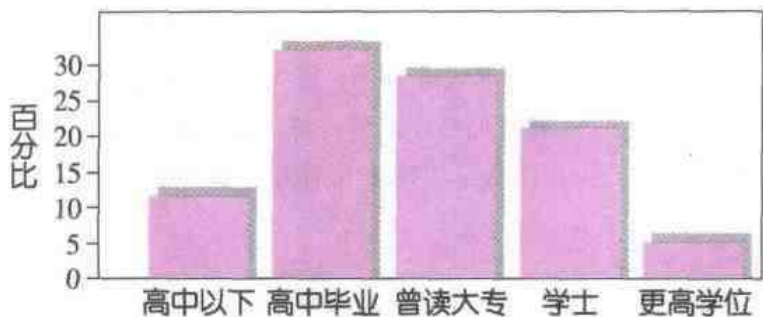


图 10.4 1998 年 25—34 岁人士教育程度分布的柱状图

当我们在考虑各种图的时候，把变量稍加分类会有帮助，有的变量具备有意义的数值尺度(numerical scale)(例如身高几厘米，SAT 分数等)，而有的变量例如性别、职业或者教育程度，只是把个体分到不同类别而已。饼状图和柱状图对于第二种变量最有用。

• 类别和数值变量

类别变量(categorical variable)把个体归类到数个组(group)或数个类别(category)其中之一。

数量变量(quantitative variable)的值是数值的，因此拿来做算术比如加法或平均的时候，是有意义的。

要表示类别变量的分布，可以用饼状图或柱状图。

虽然饼状图和柱状图都可以用来表示像教育程度这样的类别变量的分布(不论是计数还是百分比)，但柱状图的用途还是比较广些。



例3 税太高?

图 10.5 比较了八个国家的税赋高低。柱体的高度显示每个国家的国内生产总值 (GDP, gross domestic product, 意思是国内生产的所有产品及服务的总值) 被课税的百分比。习于抱怨税率太高的美国人恐怕会很惊讶的发觉: 美国的税率是 GDP 的 28.5%, 几乎是其中税赋最低的国家。

我们不能用饼状图来取代图 10.5, 因为图 10.5 比较的是 8 个个别的量, 而不是一个整体的各个部分。饼状图只能用来比较一个整体的各个部分。柱状图却可以用来比较并不属于同一个整体的数量。

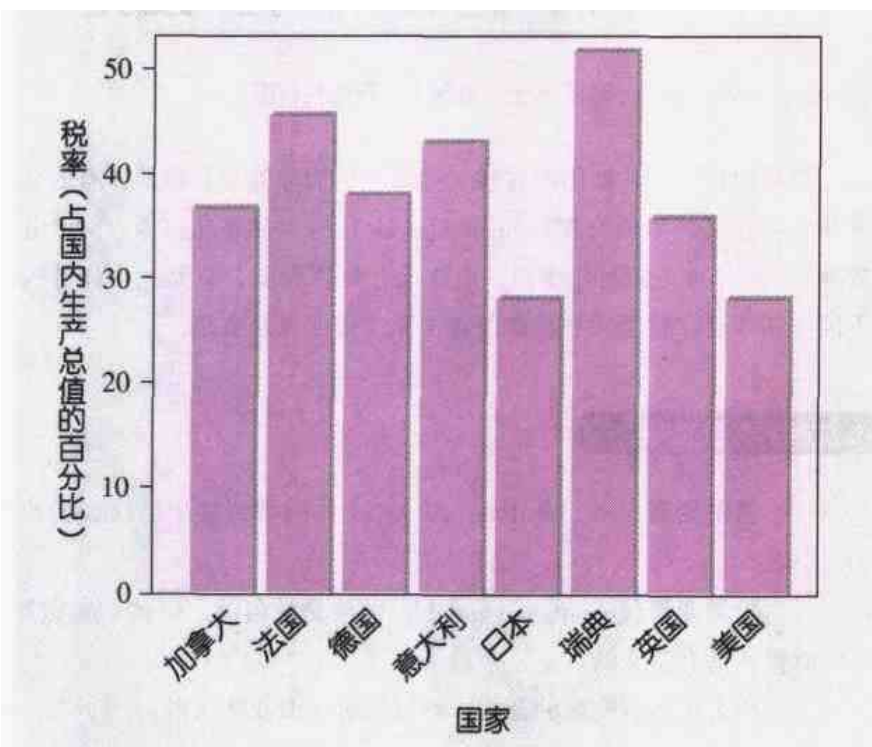


图 10.5 1996 年八个国家的税率(占国内生产总值的百分比)(资料来源: 经济合作与发展组织, 1999 年)



留意象形图

柱状图是经由比较代表各数量的柱状高度，来比较各个数量的大小。但是我们眼睛所看的，除了高度外还有面积。当所有长条的宽度都一样时，面积(宽度乘上高度)和高度成正比，所以我们的眼睛接收到的是正确的印象。当你画柱状图时，每个长条都要一样宽。要是从艺术美感的观点来看，柱状图实在有点单调，令人动念头想要用别的图取代柱状，以期更能吸引视线。

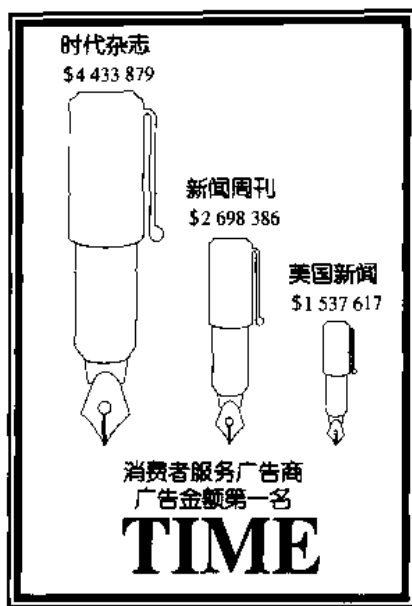


图 10.6 象形图，是柱状图的“变种”，很吸引人，但是会有误导(Copyright © 1971 by Times Inc Reproduced by permission)

例 4 会误导的图

图 10.6 是“象形图”(pictogram)。象形图其实就是柱状图，只是以图形取代柱体。

这个图的目标是广告商，他们正在考虑预算要花在什么地方。这个图显示出，《时代杂志》(Time)吸引了巨量的广告支出。真的是这样吗？笔顶端的数字显示，在《时代杂志》的广告花费是《新闻周刊》(Newsweek)的 1.64 倍。为什么从图形看来《时代杂志》的超越远不止此呢？

为了把图形放大，绘图的人必须同时把图的高和宽加大，以免变形。如果《时代杂志》笔的高和宽都是《新闻周刊》笔的 1.64 倍，面积就会是 1.64×1.64 倍，即大约 2.7 倍那么大。我们的眼睛会对笔的面积有所反应，所以就把《时代杂志》看成了大赢家。



随着时间变动的线图

许多数量变量都是隔一段时间量一次的。比如说,我们也许会度量成长中儿童的身高,或者在每个月的月底记录下某只股票的股价。在这类例子当中,我们主要感兴趣的,是变量如何随着时间变动。要表示出变量随着时间推移所产生的变化,应该使用线图(line graph)。

• 线图

一个变量的**线图**(line graph)描绘出该变量在不同的时间所量出来的结果。一定要把时间刻度放在你画的图的横轴上,而把你正在度量的变量放在纵轴上。用直线连接根据数据画出的点,以便呈现出随时间变化的情况。

例 5 汽油价格

加油站的汽油价格近年来有些什么变化?图 10.7 是美国在 20 世纪 90 年代每个月普通无铅汽油平均价格的线图。举例来说,1990 年 1 月的价格是每加仑 105.1 美分。图中的第一个点是在 1990 年年初(1 月)的正上方,和纵轴刻度 105.1 同高的位置。

如果是把每月价格列出一个长长的表,会很难看出价格走向的形态,但图 10.7 却使得这个形态清楚多了。我们要怎么看此图呢?

- 首先,找出**整体形态**。比如说,长期以来随着时间上升,或者长期下来随着时间下降,就叫**趋势**(trend)。在 20 世纪 90 年代的十年当中,美国的汽油价格并没有整体的上升趋势或整体的下降趋势。
- 其次,找找看有没有显著**偏离整体形态**的现象。1990 年价格的飙升,1997 年后期到 1998 年整年价格的持续下滑,在缺乏整体趋势的情况下显得突出。1990 年的巅峰价格,是因为伊拉克入侵科

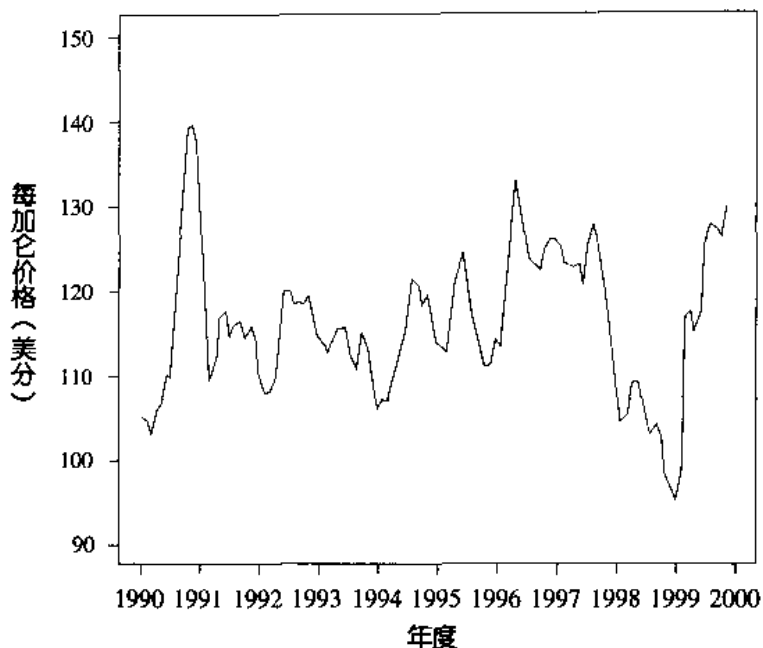


图 10.7 1990 年 1 月和 1999 年 12 月之间, 每个月普通无铅汽油平均价格的线图(数据来自美国劳工统计局)

威特。1997 和 1998 年的急速下滑, 反映出亚洲的经济危机(因此对油产品需求降低)以及汽油的供过于求, 直到石油输出国组织(OPEC)的国家靠着减产在 1999 年又把价格提升。

- 随时间变动的变量, 常常会年复一年出现有规则的季节变动(seasonal variation)现象。汽油价格通常在夏天的驾车季节最高, 而在冬天对汽油需求较少的时候最低。你可以在图 10.7 中看到, 这种夏天上升、秋天下降的现象, 在很多年当中都有出现。

因为季节变动很常见, 所以美国有许多政府统计资料都经过调整, 以便去除季节效应。举例来说, 因为圣诞节的销售工作结束以及北方的户外工作由于冬季气候而减少, 所以每年在 1 月份时美国的失业率会升高。如果政府公布的失业率每到 1 月就跃升, 可能会造成困扰(还可能有政治上的麻烦)。美国劳工统计局知道每年 1 月失业率大约要上升多少, 所以依据预期中的改变调整公布的资料。只有实际的失业率上升得比预期还多时, 公布的失业率才会上升。这样我们才能看到就业情况的实质改变, 而不会让固定的季节变动给弄糊涂了。

越南效应

民间传说在越战期间许多美国人为了规避服役而去上大学。统计学家韦纳(Howard Wainer)用数据对应时间画了图, 试图找出关于“越南效应”的蛛丝马迹。他发现入伍资格测验(Armed Forces Qualifying Test 是给新兵做的智力测验)的分数在越战期间急剧降低, 然后又回升。申请入大学的学生要考的 SAT 测验的分数, 在战争开始初期也同样下降。似乎选择大学而不愿从军的人, 把两个测验的平均分数都降低了。



● 季节变动, 季节调整

在已知的固定间隔时间重复的形态, 叫做**季节变动**(seasonal variation)。而许多在固定间隔时间度量的资料, 都经过**季节调整**(seasonal adjustment), 也就是说, 在资料公布之前, 预期的季节变动已先消除。

注意刻度

因为图给人的印象很深刻, 所以不小心的人很容易被误导。谨慎的人在读线图时, 会很仔细地看横轴和纵轴上标示的刻度。

例 6 同居

未婚同居的人数, 近年来增加很多, 以至于有些人认为, 同居已延缓甚至取代了婚姻。图 10.8 里是美国未婚同居住户数的两个线图。数据又是出自当前人口调查。左边的图显示出稳定而幅度不大的成长。右边的图却告诉我们, 同居人数正在激增。

其中的奥秘在于刻度。要把左边的图变成右边的图, 只要把纵轴拉长, 横轴压缩, 然后把线图两头未到达的纵轴刻度切掉。现在你已经知道如果要夸大或者压低一个线图的上升或下降趋势, 应该怎么做了。

这两个图哪个是正确的呢? 两个都是对应数据的正确图形, 但是两者都对刻度做了选择, 以便制造出特定的效果。线图并没有所谓的“正确”刻度, 通过对刻度的选择, 同样都是正确的图形也可以给人很不同的印象。所以要小心刻度!

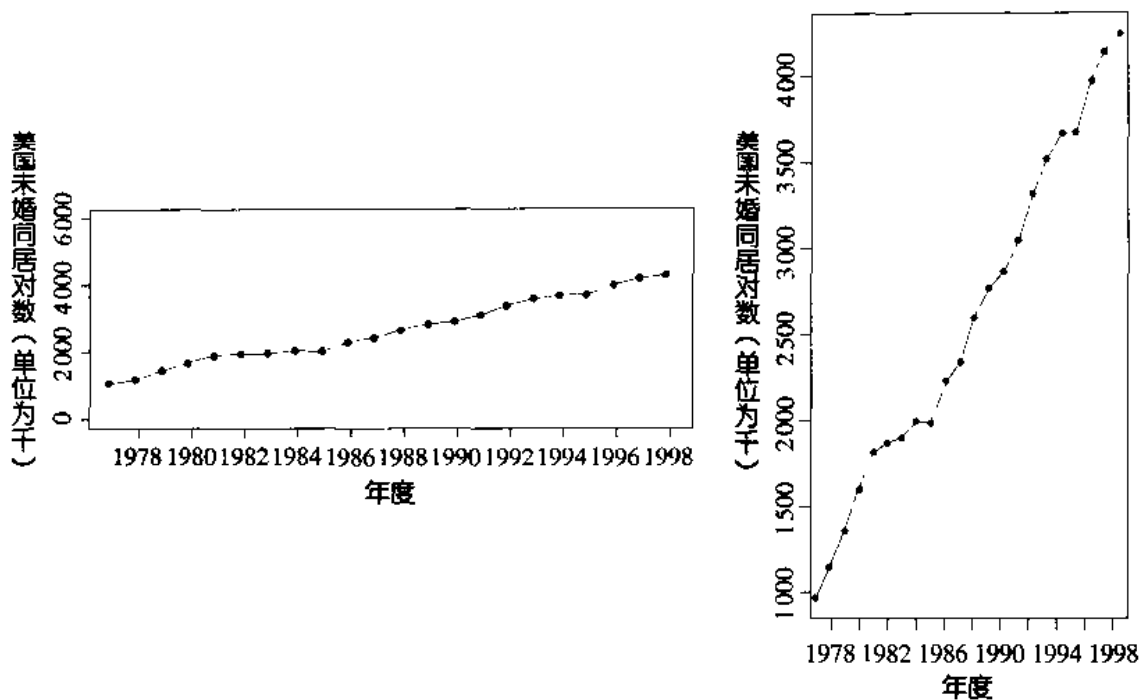


图 10.8 改变线图的刻度所产生的效果。两个图所展示的是同一组数据，但是右边的图使得增加的速度看起来快得多

怎样把图画好

图画是传达数据信息最有效的方式。好的图常常可以把数据当中的信息清楚显示出来，而若只把数据制成表，可能很难甚至不可能做到这点。还不止这样，比起数值资料所制造的印象，图所制造的直观视觉效果强多了。以下是把图画好的一些原则。

- 一定要在标示和说明里表示清楚，图里面画的变量是什么，单位是什么，以及资料来源。
- 要让数据很醒目。要确实注意到，抓住看图者注意力的是数据本身，而不是标示、格子，也不是背景的图样。你是在画一个呈现数据的图，不是在从事艺术创作。
- 要注意实际上眼睛会捕捉到什么。避免用象形图，而且要小心选择刻度。也不要很炫目的“三维空间”效果，因为那只会让人看得迷迷糊糊，不会增加看的人对数据的了解。考虑一下是不是把图稍微做些改变，使信息更清楚。

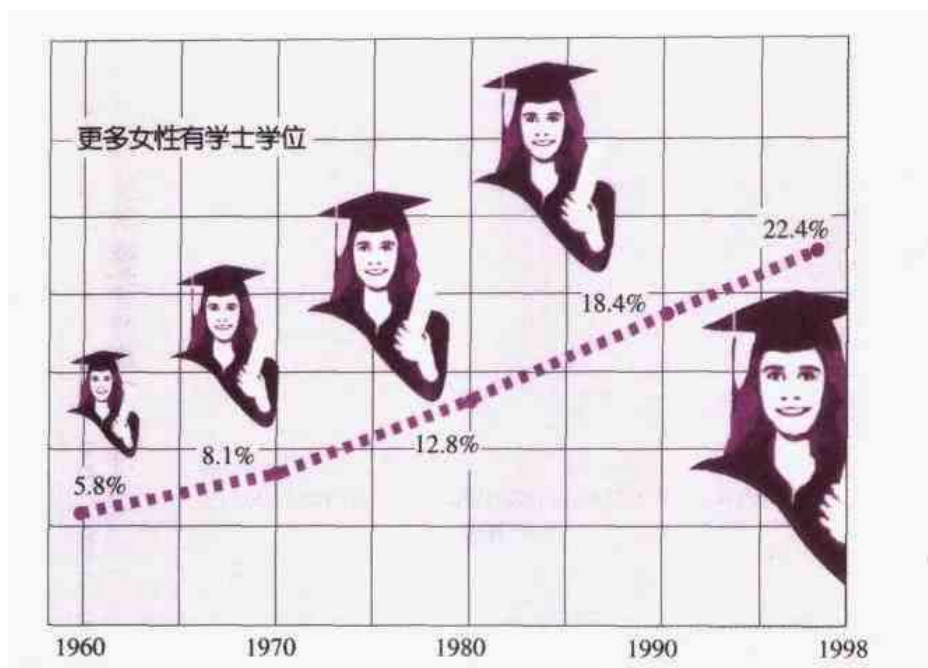


图 10.9 蹩脚图：这个图花费了许多不必要的笔墨，反倒不易看清数据

例 7 受高等教育的人增加了

图 10.9 中显示出，25 岁以上女性拥有学士或更高学位的比率增加了。一共只有 5 个数据点(data point)，所以线图应该很简单。但图 10.9 可不简单，画图的人大概忍不住要在背景当中加画些东西，又在图上加了格子线。这样子反倒比较看不清楚数据了。图上面的格子线毫无用处，因为如果你的观众必须知道确实的数字，可在图外再列一个表。好的图目的在清楚呈现资料，不需浪费笔墨去画蛇添足。

例 8 税太高，续集

图 10.5 是一个很清楚的柱状图，比较了 8 个国家的税率(GDP 的百分比)。国家的顺序是照国名的英文字母顺序排的。如果按照税率高低排，会不会更清楚呢？

图10.10 就这样做了。这个小小的改变把图改善了，现在很容易看出，每一个国家的税率和其他国家比起来，会排在什么位置。

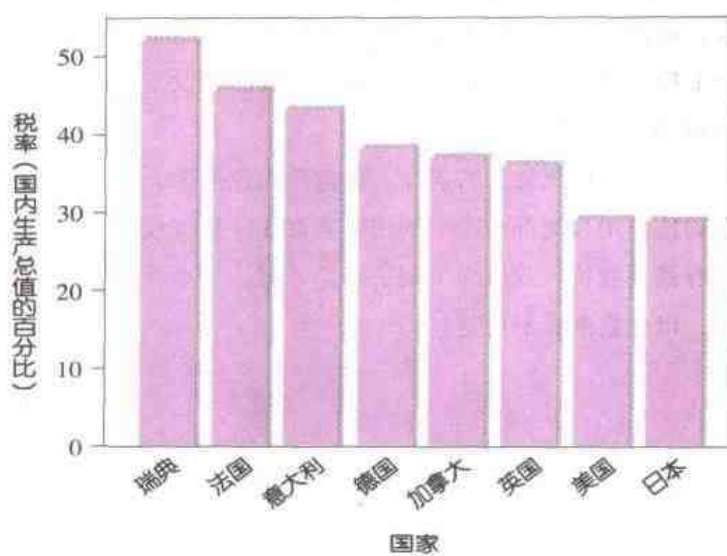


图 10.10 1996 年八个国家的税率(占国内生产总值的百分比)。把柱体的顺序改变之后，图 10.5 的图就变得更完善了



本章重点摘要

想知道数据说了什么，就先画个图。要画怎样的图，要看数据的类型而定。你的变量是不是像教育程度或者职业之类，只是把每个个体归类的**类别变量**？还是用有意义的数值单位度量的**数量变量**？要展示类别变量的分布，得用**饼状图**或**柱体图**。饼状图永远是显示整体的各个部分，而柱体图可以用来比较任何用同样单位度量出来的数。要表示出一个数量变量如何随着时间改变就用**线图**，以变量的值(纵轴)对应着时间(横轴)画图。

图也可能造成视觉上的误导。要避免用**象形图**，因为这种图把长线图里的柱体用长宽都会改变的图画来取代。检查线图里的刻度，看看有没有被刻意拉大或压缩来制造特定效果。不要在图中加入不必要的东西，以免数据看不清楚。



第 10 章 习题

10.1 彩券销售。美国各州都有很多种彩券。表 10.2 中呈现出各种彩券的销售金额。用柱体图表示出不同彩券销售金额的分布情况。这些资料用饼状图来表示是否合适?

表 10.2 不同种类的州彩券销售金额, 1998 年

销售金额(以百万美元为单位)	
即时彩券	13 882
3 位数彩券	5 643
4 位数彩券	2 232
乐透彩券	9 854
其他金额	3 978
总计	35 588

资料来源:《1999 年美国统计精粹》

10.2 数字相符吗?从表 10.2 里可以看得到,美国各州彩券的销售总金额中,各种彩券的个别贡献各有多少。花在 5 种彩券上的金额,总和是多少?为什么这个总和与表里面的总金额不完全相符?

10.3 婚姻状况。在《美国统计精粹》里面,有关 1998 年美国成年女性婚姻状况中,可以找到下面这些资料:

婚姻状况	计数(千人)
未婚	21 043
已婚	59 255
寡居	11 027
离婚	11 078

(a) 在 1998 年,有多少女性是没有配偶的?

(b) 用柱体图来表示出婚姻状况的分布。



(c) 也可以用饼状图来表示吗?

10.4 我们的利率比较高。图 10.11 是从一个投资公司的广告中选出来的图, 该公司保证, 他们的利率会高过银行的户头以及其他投资途径。这个图能否正确的比较标示出来的 4 种利率? 请说明。

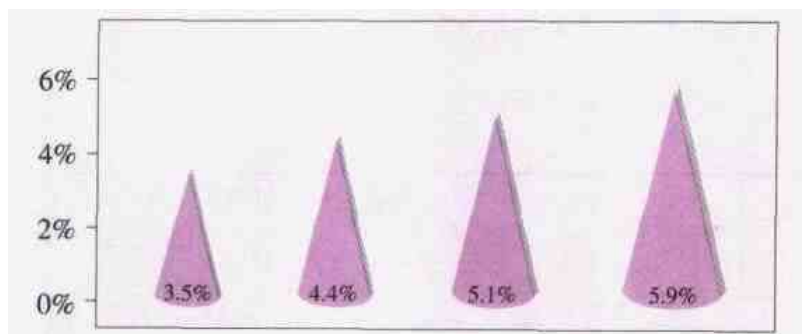


图 10.11 利率的比较。对照习题 10.4

10.5 我们卖 CD。在听歌的人开始下载歌曲而不再买 CD 之前, 哪家从业者在网上卖出最多 CD? 图 10.12 画出了 1997 年业绩领先的几家, 以及他们在网上 CD 销售总额中所占的比率。用这样的图来代表它们对应的那些数据, 公不公平? 请说明你的看法。



图 10.12 1997 年 CD 网上销售前三名的从业者, 及其销售业绩占总销售额的百分比。对照习题 10.5 (资料来源: Jupiter Communications)

10.6 谋杀案的凶器。在 1999 年的《美国统计精粹》里, 有美国联邦调查局对 1997 年的谋杀案相关数据的报告。在那一年的所有谋杀案当中, 有 53.3% 用手枪当凶器, 14.5% 用其他枪支, 13.0% 用



刀, 6.3%用身体的某部分(通常是手或脚), 还有 4.6%用钝器。画一个图来显示这组资料。你需不需要用一个“其他方法”的类别?

10.7 新鲜橘子的价格。图 10.13 是从 1990 年 1 月到 2000 年 1 月之间, 每个月新鲜橘子平均价格的线图。资料来自劳工统计局的每月零售价格调查, 而纵轴上的数字, 是价格指数(index numbers), 而不是实际的价钱。价格指数是把每个月的价格, 表示成某一个基期(base period)价格的百分比, 这个例子中所指的基期是 1982—1984 年。所以 150 的意思: “基期价格的 150%”。

- (a) 图中显示出很强的季节变动。怎么样从图里看出这一点? 你觉得为什么橘价应该会有季节变动?
- (b) 在把季节变动纳入考虑后, 这段期间橘价的整体趋势如何?

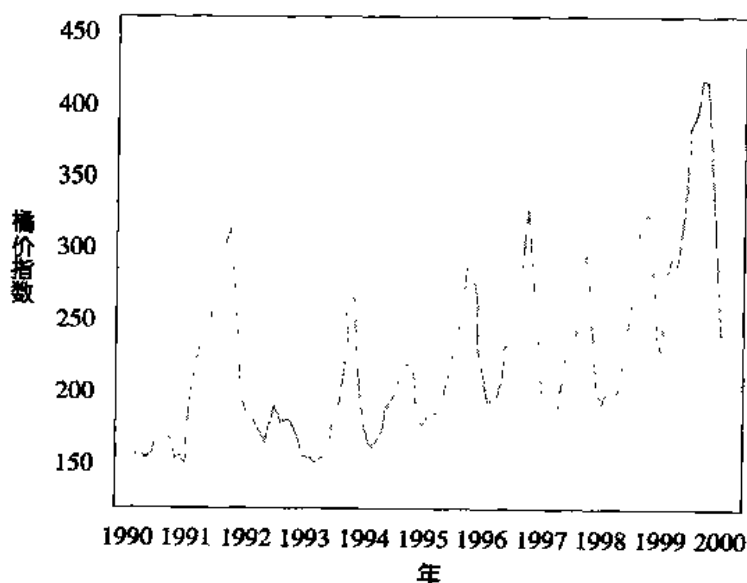


图 10.13 1990 年 1 月—2000 年 1 月的新鲜橘子价格, 对照习题 10.7

10.8 大学新生。1998 年一份对大学新生做的调查, 询问了他们准备专攻什么领域。结果如后: 10% 艺术及人文学科, 17% 商业, 11% 教育, 16% 理工, 15% 专业领域及 8% 社会学科。(资料来自 1999 年的《美国统计精粹》。)

- (a) 准备要读上面未列出的领域的学生, 占多少百分比?
- (b) 画一个图来比较大学新生计划读各种不同领域的百分比。



10.9 出口. 图 10.4 比较了 1997 年世界主要出口国家: 德国、日本及美国的出口金额(美元)。

(a) 说明一下为什么这个图不正确。

(b) 画一个正确的图来比较这三个国家的出口金额。

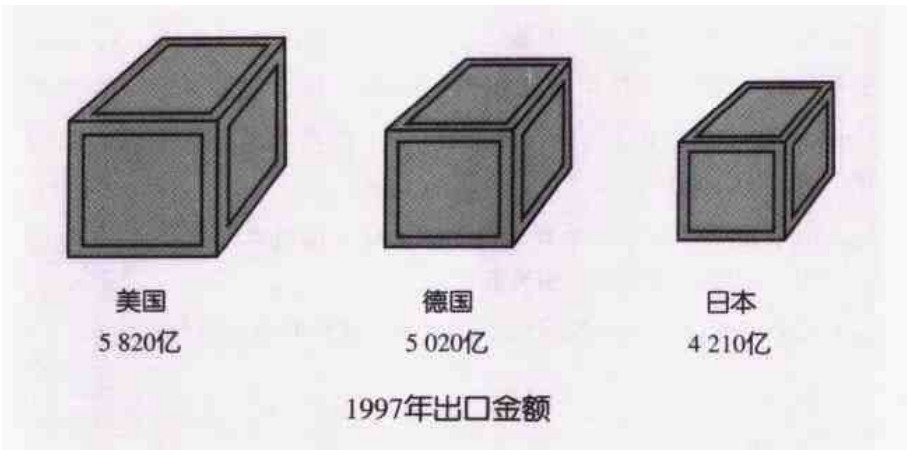


图 10.14 德国、日本及美国的 1997 年出口金额。对照习题 10.9(资料来源: 经济合作与发展组织)

10.10 老百姓骚乱事件. 在 1970 年上下几年, 美国许多大城市都发生过骚乱。以下的政府资料, 显现出 1968—1972 年之间, 每 3 个月发生的骚乱事件的次数。

(a) 针对这组资料画一个线图。

(b) 资料显示出长期趋势, 以及一年当中的季节变动。形容一下是怎

期间	次数	期间	次数
1968 年 1 月—3 月	6	1970 年 7 月—9 月	20
4 月—6 月	46	10 月—12 月	6
7 月—9 月	25	1971 年 1 月—3 月	12
10 月—12 月	3	4 月—6 月	21
1969 年 1 月—3 月	5	7 月—9 月	5
4 月—6 月	27	10 月—12 月	1
7 月—9 月	19	1972 年 1 月—3 月	3
10 月—12 月	6	4 月—6 月	8
1970 年 1 月—3 月	26	7 月—9 月	5
4 月—6 月	24	10 月—12 月	5



样的趋势和怎样的季节变动。你对这些骚动事件的季节变动，有什么解释？

10.11 未婚生育。以下是未婚生育占全美生育百分比的资料，出自《美国统计精粹》。数字明显的随时间而上升。根据这些数据画两个线图一个图显示出缓慢上升，而第二个图显示出惊人的增长。

年度	1960	1965	1970	1975	1980	1985	1990	1995
未婚生育所占百分比	5.3	7.7	10.7	14.2	18.4	22.0	28.0	32.2

10.12 此图糟不糟？图 10.15，出现在 1975 年 10 月 5 日肯塔基州列克星顿的《前锋报》(Herald-Leader)。讨论一下这个图是否正确。

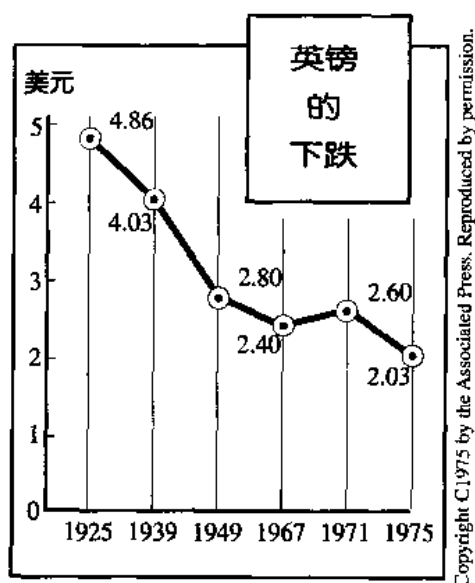


图 10.15 出现在报纸上的英镑价值图，对照习题 10.12

10.13 趋势。以下这些资料中，有哪些你认为会显示出趋势？显现出的趋势会是上升还是下降？(资料是每年记录的。)

- (a) 进大学的学生中会携带打字机的百分比。
- (b) 进大学的学生中会携带个人电脑的百分比。
- (c) 成年女性没有在家庭以外的地方工作的百分比。

10.14 季节变动。假设你每个月检视芝加哥的平均温度，持续很多



年。你觉得这组资料的线图会出现季节变动吗?描述一下你预期见到怎样的季节变动。

10.15 销售增加了。你新开的礼品店在12月的销售额是11月的两倍。你应不应该下结论,认为你的店生意愈来愈好,你就快要发财了?答案要加以说明。

10.16 计算有多少人失业,新闻报道说:“6月份在美国有工作的人,比1990年底以来任何一个月份都要多。如果你觉得这和上礼拜你在报纸上读到的不一样,不用担心。的确是不一样。那份报告说就业人口在6月份大减,非农业就业人数共减少了117 000人,靠这消息帮忙,美国联邦储备委员会(Federal Reserve)又降息了……然而事实上6月份的就业人数比5月份增加了457 000人。”就业人数实际上增加了457 000人,而官方的报告却说是减少了117 000,这中间的差距应该如何解释?

10.17 太阳黑子的周期。对应时间画图,可以画出高低起伏的周期性活动。图10.16是20世纪当中每一个月,太阳朝地球的这面图子

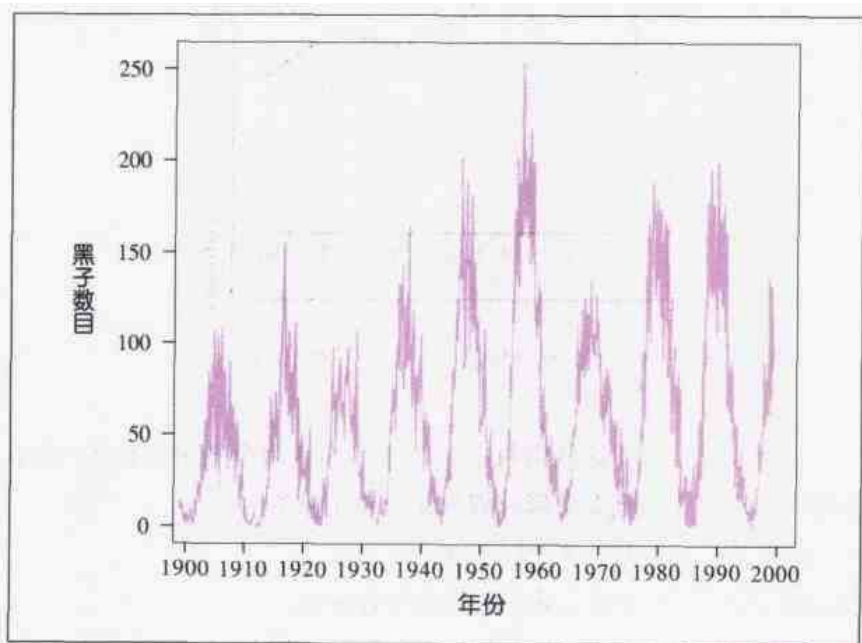


图10.16 黑子周期,对照习题10.17。这是1900—1999年之间每个月黑子数目的线图(资料来源:美国全国海洋和大气署, National Oceanic and Atmospheric Administration.)



平均数目的线图。黑子周期大约是多长?也就是说,图里面连续两个“波谷”大约隔几年,且黑子数目的消长,有没有整体趋势?

10.18 卡车、小客车大对决。消费者开始用卡车、SUV 及小厢型车来取代小客车。以下有美国小客车和卡车的新车销售资料。

这里列的卡车包括了 SUV 和小厢型车。在同样的轴上画两个线图,来比较小客车和卡车销售随时间改变的情况。描述一下你看到的趋势。

年份	1981	1983	1985	1987	1989
小客车(千辆)	8 536	9 182	11 042	10 277	9 772
卡车(千辆)	2 260	3 129	4 682	4 912	4 941

年份	1991	1993	1995	1997	1999
小客车(千辆)	8 175	8 518	8 636	8 273	8 697
卡车(千辆)	4 365	5 681	6 481	7 226	8 717

10.19 谁在卖小客车?图 10.17 是 1997 年各制造商小客车销售额百分比的饼状图。绘图者为了让图较吸引人,用了一个车轮来当作饼状图的圆。这样的图是不是忠实呈现了资料?请说明。

10.20 谁在卖小客车?对应习题 10.19 的资料,画一张柱体图。与图 10.17 的饼状图相比,你的柱体图有些什么优点?

10.21 边境巡逻队。以下数字是 1971—1997 年之间,美国边境巡逻队抓到的应驱逐出境的外国人人数。用图表来展示这些数据。从这些数据看出来的最重要信息是什么?

年份	1971	1973	1975	1977	1979	1981	1983
计数(千人)	420	656	767	1 042	1 076	976	1 251

年份	1985	1987	1989	1991	1993	1995	1997
计数(千人)	1 349	1 190	954	1 198	1 327	1 395	1 536

10.22 照着以下所规定的特征,分别画出线图。在代表时间的轴上标示出年度。

- (a) 有很强的下降趋势,但是没有季节变动。
- (b) 每一年里都有季节变动,但是没有明显的趋势。

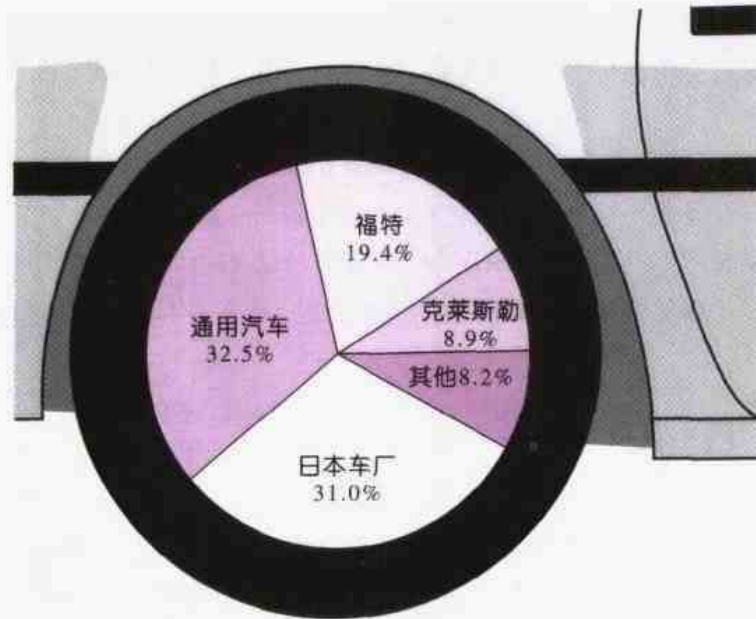


图 10.17 1997 年各制造商的小客车销售百分比，对照习题 10.19

(c) 有很强的下降趋势，加上每年的季节变动。

10.23 1 美元能买多少东西? 1 美元的购买力随着时间改变。劳工统计局借着度量属于“市场总览”(market basket)的商品和服务的总价，来算出消费者物价指数(CPI, Consumer Price Index)。如果 CPI 是 120，则在基期(base period)要花费 100 美元的商品和服务，现在要花费 120 美元。下面的数据是 1970—1998 年之间，每隔两年的年度平均 CPI。基期是从 1982—1984 年。

- (a) 画一个图来显示出，CPI 如何随着时间改变。
- (b) 这段期间价格的整体趋势是什么? 有没有哪几年是和趋势相反的?
- (c) 哪几年价格上涨最快? 哪一段时期上涨最慢?

年份	CPI	年份	CPI	年份	CPI
1970	38.8	1980	82.4	1990	130.7
1972	41.8	1982	96.5	1992	140.3
1974	49.3	1984	103.9	1994	148.2
1976	56.9	1986	109.6	1996	156.9
1978	65.2	1988	118.3	1998	163.0

10.24 坏习惯。根据《全国药物滥用家庭调查》，年龄在 12—17 岁



之间的青少年有 20.5% 在 1997 年曾饮酒, 9.4% 吸食大麻, 1.0% 吸食可卡因, 以及 19.9% 抽烟。说明为什么用饼状图呈现这些资料是不正确的。

10.25 意外死亡。美国在 1997 年共有 92 353 人意外死亡。其中有 42 340 人死于车祸, 11 858 人死于坠落, 10 163 人死于中毒, 4 051 人溺毙以及 3 601 人死于火灾。

- (a) 把每种死因的意外死亡百分比算出来, 四舍五入到整数的百分比。意外死亡中有多少百分比是属于其他死因?
- (b) 以意外死亡死因的分布画图, 标示要清楚。

10.26 货币市场基金的利润。很多人投资货币市场基金。这是一种共同基金, 其价格会试图维持在每股 1 美元, 然而会每月付利息。表 10.3 列出了自 1973 年以来, 所有可课税的货币市场基金付出的平均年利率(百分点), 1973 是有这种基金之后的第一个完整年。

- (a) 对应表中资料, 为这些年来货币市场基金所付出的利息画一个线图。
- (b) 利率像许多经济变量一样, 会有循环(cycles), 这是指很清楚但不规则的上下变动。在哪几年利率循环到达暂时的高峰?

年	利率	年	利率	年	利率	年	利率
1973	7.60	1979	10.92	1985	7.77	1991	5.70
1974	10.79	1980	12.88	1986	6.30	1992	3.31
1975	6.39	1981	17.16	1987	6.17	1993	2.62
1976	5.11	1982	12.55	1988	7.09	1994	3.65
1977	4.92	1983	8.69	1989	8.85	1995	5.37
1978	7.25	1984	10.21	1990	7.81	1996	4.80

表 10.3 货币市场基金的平均利息, 1973—1996 年

10.27 波士顿马拉松。波士顿马拉松从 1972 年开始准许女选手参赛。从 1972 年到 2000 年的女子冠军成绩列在表 10.4 里面(成绩以分钟计, 经过四舍五人)。

- (a) 为这些得胜成绩画一个线图。
- (b) 稍微描述一下, 这些年来波士顿马拉松的女子冠军成绩有怎样的



形态?最近几年是否成绩不再进步?

表 10.4 波士顿马拉松赛女子冠军成绩, 1972—2000 年

年	时间	年	时间	年	时间
1972	190	1982	150	1992	144
1973	186	1983	143	1993	145
1974	167	1984	149	1994	142
1975	162	1985	154	1995	145
1976	167	1986	145	1996	147
1977	168	1987	146	1997	146
1978	165	1988	145	1998	143
1979	155	1989	144	1999	143
1980	154	1990	145	2000	146
1981	147	1991	144		

第 11 章

用图形呈现分布

我的学校排名第几？

唐亚是她的家乡密歇根州山谷州立大学的学生。山谷州大收本州学生 4 108 美元的学杂费。唐亚想要知道，山谷州大的学费和其他密歇根州的大专院校比较起来算不算贵。

密歇根州一共有 81 所大专院校。这些学校 1999 学年的学杂费，从卡拉马助谷社区学院的 1 260 美元，到卡拉马助学院的 19 258 美元。要把山谷州大和 80 所其他学校比较不太容易，我们就来画一个图。从图 11.1 的直方图可以看出，很多所大专收的费用介于 18 000—20 000 美元之间。山谷州大是收费在 4 000—6 000 美元之间的五所学校之一。

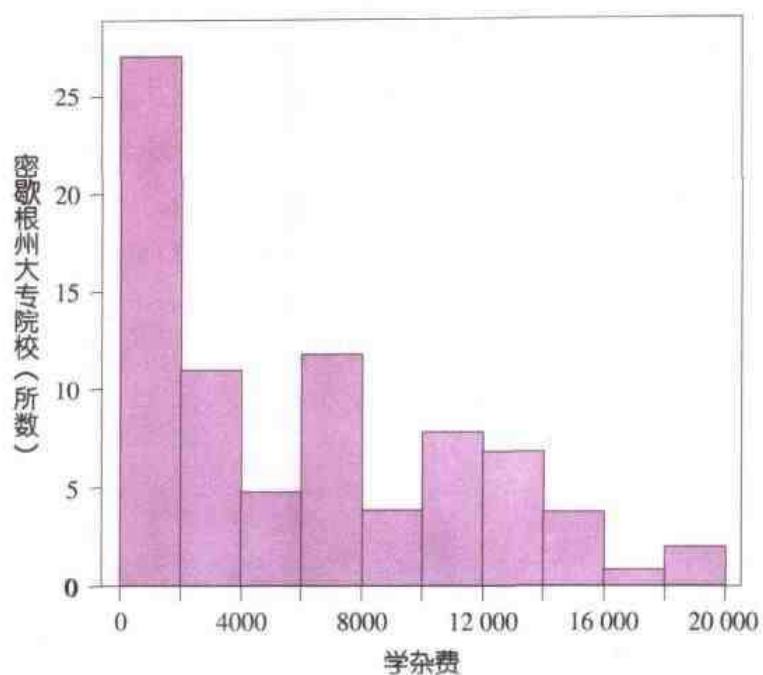


图 11.1 1999 学年密歇根州 81 所大专院校所收学杂费的直方图

1	3 5 5 5 5 5 5 5 6 6 6 6 6 6 7 7 7 7 7 8 8 8 8 9 9 9
2	8
3	1 5 5 6 6 6 8 9 9
4	0 1 2 5
5	0 1
6	1 3 3 4 6 6
7	0 1 4 5 6 7
8	1 6 9
9	7
10	1 3 9
11	1 3 6 7 8
12	3 9
13	2 4 4 5 7
14	3 9
15	1 6
16	0
17	
18	2
19	3

图 11.2 密歇根州学杂费资料的茎叶图



表 11.1 各州 65 岁以上居民百分比(1998 年)

州	百分比(%)	州	百分比(%)	州	百分比(%)
亚拉巴马	13.1	路易斯安那	11.5	俄亥俄	13.4
阿拉斯加	5.5	缅因	14.1	俄克拉何马	13.4
阿里桑纳	13.2	马里兰	11.5	俄勒冈	13.2
阿肯色	14.3	马萨诸塞	14.0	宾州	15.9
加州	11.1	密歇根	12.5	罗得岛	15.6
科罗拉多	10.1	明尼苏达	12.3	南卡罗来纳	12.2
康涅狄格	14.3	密西西比	12.2	南达科他	14.3
特拉华	13.0	密苏里	13.7	田纳西	12.5
佛罗里达	18.3	蒙大拿	13.3	德州	10.1
乔治亚	9.9	内布拉斯加	13.8	犹他	8.8
夏威夷	13.3	内华达	11.5	佛蒙特	12.3
爱达荷	11.3	新罕布什尔	12.0	弗吉尼亚	11.3
伊利诺伊	12.4	新泽西	13.6	华盛顿	11.5
印第安纳	12.5	新墨西哥	11.4	西弗吉尼亚	15.2
艾奥瓦	15.1	纽约	13.3	威斯康星	13.2
堪萨斯	13.5	北卡罗来纳	12.5	怀俄明	11.5
肯塔基	12.5	北达科他	14.4		

$$5.0 \leq 65 \text{ 岁以上居民比率} < 6.0$$

$$6.0 \leq 65 \text{ 岁以上居民比率} < 7.0$$

$$18.0 \leq 65 \text{ 岁以上居民比率} < 19.0$$

一定要把组界定义得非常明确, 每个个体只能被归入一个组。若某一州 65 岁以上的居民占全州人口的 5.9%, 则这一州属于第一组, 如果是 6.0% 就属于第二组。

第 2 步: 数一下每组中每个个体的个数, 以下就是数出来的结果。

组间	计数	组间	计数	组间	计数
5.0—5.9	1	10.0—10.9	2	15.0—15.9	4
6.0—6.9	0	11.0—11.9	9	16.0—16.9	0
7.0—7.9	0	12.0—12.9	11	17.0—17.9	1
8.0—8.9	1	13.0—13.9	14	18.0—18.9	1
9.0—9.9	1	14.0—14.9	6		



第3步：画直方图。把要展示分布的变量在横轴上标示出刻度。在这个例子当中，这个变量就是“65岁以上居民所占比率”。刻度从4到20，这样就包含了我们选定的组的所有范围。然后把计数的刻度标示在纵轴上。每一个柱体代表一组，柱体底部涵盖该组的范围，而柱体的高度代表该组的计数。柱体与柱体之间不要有空隙，除非有一组是空的，此时它对应的柱体高度是零。图11.3就是我们的直方图。

就跟柱体图一样，我们的眼睛会对直方图的柱体面积起反应，因此要确定直方图每个组的宽度一样。如何分组并没有绝对的标准。不过如果组数太少，会组成“摩天楼”直方图，所有的值只落在少数几个组里面，而那几个组的柱体会很高；而分了太多的组，又会造成“煎饼”图，大部分的组只有一个观测值，甚至没有观测值，到处平平的。这两种选择都不能有效描绘出分布的形状。要展示出形状，你得自己判断怎样选择合适的组数，也有统计软件会帮你决定怎样分组。电脑的选择通常很不错，不过你要修改也可以。

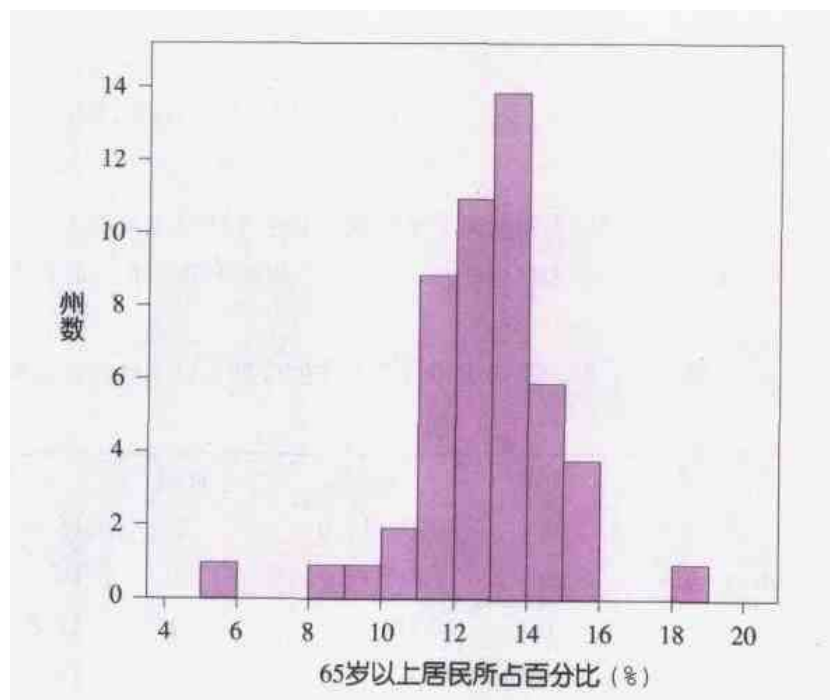


图11.3 美国50个州中，65岁以上居民占各州人口百分比的直方图。注意有两个异常值



解释直方图

画统计图本身并不是最终目的，画统计图的目的是要帮助我们了解资料。在你(或你的电脑)画完图之后，一定要问：“我看到了什么？”以下是审视图形的一般策略：

• 形态与偏差

在任何一组资料的图形里，我们要找的是一般形态(overall pattern)，以及有异于一般形态的显著偏差(deviation)。

这个策略我们已经在线图上应用过。趋势和季节变动都是线图中常见的一般形态。图 10.7 中 1990 年因伊拉克入侵科威特所造成的油价飙升，就是有异于一般形态的偏差之例。以图 11.3 的直方图而言，从直方图中异于一般形态的偏差说起，会更容易些。有两个州与众不同，一旦从直方图中注意到这两个州，就可以到表里面去查出是哪两个州。佛罗里达州有 18.3% 的居民是在 65 岁以上，而阿拉斯加则只有 5.5%。这两州是很明显的异常值。

• 异常值

一组资料的任何图形的异常值(outlier)，是指落在图形一般形态之外的观测值。

犹他州的 65 岁以上居民占 8.8%，算不算是异常值呢？某个观测值到底算不算是异常值，在某种程度上是主观判断的问题。犹他州是观测值主体群中最小的一个，并没有像佛罗里达和阿拉斯加那样脱离一般形态。我不会叫它异常值。一旦你找到异常值，就应该寻求解释。许多异常值其实是错误造成的，比如把 4.0 打成 40。有些异常值则显示出某些观测值的特性。要解释异常值，通常需要些背景知识。佛罗里达有许多退休人口，所以 65 岁以上比率高，一点儿也不奇怪；而阿拉斯加在北部边陲地带，年纪大的人少，也很正



常。

要找出直方图的一般形态，得先把异常值撇开不看。我们可以用以下的简单方法来做考虑。

• 分布的一般形态

要描述分布的一般形态：

- 找出**中心**(center)及**离度**(spread)。检查看看该分布是否有简单的**形状**(shape)，可以很容易描述。

在第 12 章里我们会学到怎样用数值来描述中心和离度。目前我们不防就用分布的中间点(midpoint)来表示分布的中心，中间点就是差不多有一半观测值比它小，一半比它大的那个点。也可以不考虑异常值，只用最小和最大的值，来描述分布的离度。

例 2 描述分布

再来看看图 11.3 当中的直方图。**形状**：这个分布只有一个尖峰(peak)。它大致对称(symmetric)，也就是说图的形态在尖峰的两边很相似。**中心**：分布的中间点很接近尖峰的位置，即在 13% 附近。**离度**：离度差不多是从 9% 到 16%，如果我们不计入两个异常值的话。

密西根州大专院校学杂费的分布，如图 11.1 所示，**形状**就很不一样了。在收费最低的组那边，有很突出的尖峰。大部分大专院校收的钱低于 8 000 美元，但是右边有一条长尾巴，一直延伸到接近 20 000 美元。我们把一端有一条长尾巴的分布称做偏斜的(skewed)。**中心**大约是在 4 500 美元(有一半学校收的钱比这个少)。**离度**很大，从 1 260 美元到超过 19 000 美元。没有异常值：学费最高的几所学校只不过是长尾巴的延伸，属于一般形态。

要描述一个分布的时候，注意力要放对重点。要寻找主要的尖峰，而不是直方图中的小起伏，例如像图 11.1 里面那些。要辨别出



明确的异常值，而不是直接把最小和最大的观测值就当做异常值。还要看看是否大致有对称性，还是有明显的偏斜。

• 对称及偏斜分布

若直方图的左半和右半大致上可看成互为镜中影像，则称该分布为**对称**(symmetric)。

假如直方图的右边(包含较大观测值的那一半)延伸出去比左边远得多，则这个分布是**右偏**(skewed to the right)。假如直方图的左边延伸出去比右边远很多，称这个分布是**左偏**(skewed to the left)。

以数学的定义来说，对称的意思是说：一个图(比如直方图)左右两半确实是互为镜中影像。资料几乎不会完全对称，因此我们愿意把像图 11.3 的直方图叫做大致对称，做为整体描述。然而像图 11.1 里的学费分布，却是明显右偏。以下还有其他例子。

例 3 又见抽样

从同一总体抽许多随机样本，所得到统计量的值，会形成有规律形态的分布。图 11.4 的直方图展示的是我们在第 3 章见过的分布。抽取 1 523 位成人的简单随机样本，问其中每一个人，在过去 12 个月中有没有买过彩券。回答“有”的比例，就是样本比例 \hat{p} 。同样的步骤做 1 000 遍，可以从这 1 000 个随机样本得到 1 000 个样本比例 \hat{p} 。图 11.4 显示的，是在真正情况是总体中有 60% 的人曾购买彩券的假设下，1 000 个样本比例的分布状况。

这个分布大致对中间部位的单一尖峰对称。中心点在 0.60，反映出统计量的无偏性质。1 000 个值从最小到最大的离度，是 0.556—0.643。

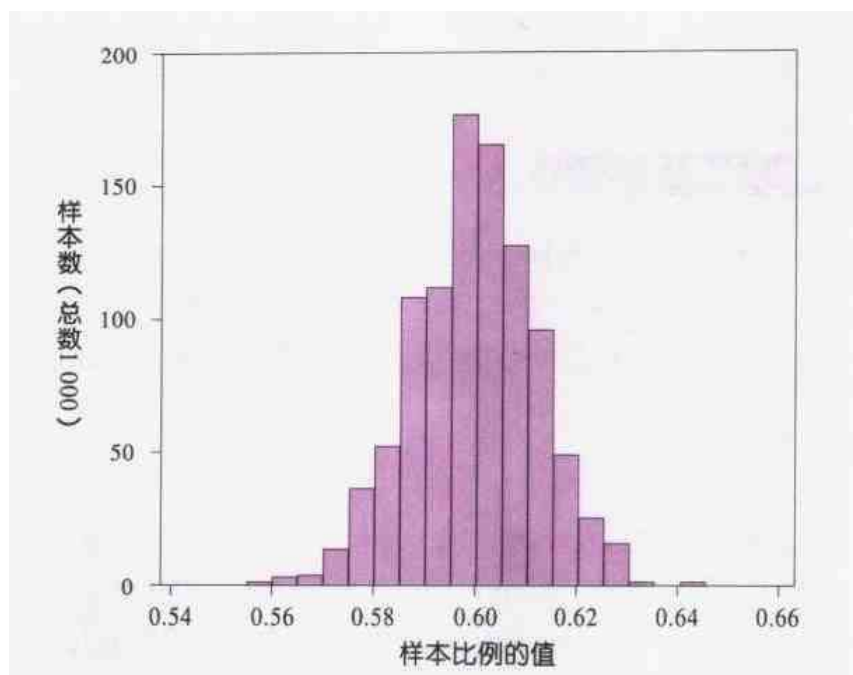


图 11.4 从同一总体抽取 1 000 个简单随机样本, 所得样本比例 \hat{p} 的直方图。这个分布是对称的

例 4 莎士比亚用字

图 11.5 显示莎士比亚(Shakespeare)的戏剧中用字长度的分布。这个分布有一个尖峰, 而且大致右偏, 剧中用了许多短字(3 到 4 个字母)及少数很长的字(10、11 或 12 个字母), 使得直方图的右尾延伸得比左尾远。分布的中心约略是 4。也就是说, 莎士比亚的用字约有一半是 4 个或更少字母。离度是从 1 个字母到 12 个字母。

请注意: 图 11.5 的纵轴并不是字数, 而是莎翁用字中各种长度所占百分比。当计数很大时, 或者当我们想比较数个分布时, 用比例做直方图要比用计数来做更方便。不同的写作风格会有不一样的字长分布, 但是都会右偏, 因为短的字很常见, 而很长的字则比较稀少。

分布的整体形状, 提供了关于变量的重要信息。有些类型的资料, 总是会造就出对称分布, 有些又总是造就出偏斜分布。例如同一

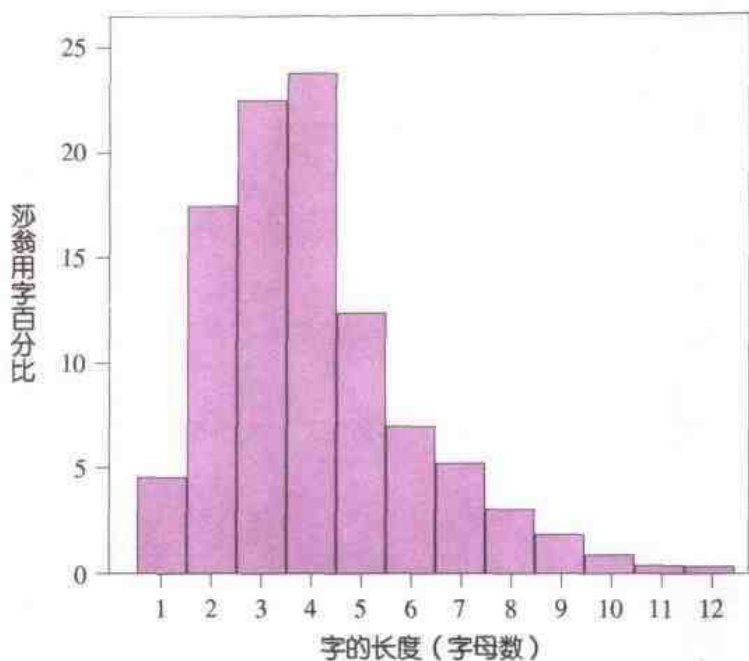


图 11.5 莎士比亚剧中用字长度的分布。这个分布是右偏的

种生物的大小(比如蟋蟀的长度),就常常是对称的。而收入的资料(不管是个人的、公司的,还是全国的),通常明显右偏。普通收入的有很多,高收入的有一些,然后有极少数的超高收入。不过要记得也有很多分布的形状是既不对称又不偏斜的。有些数据会显出其他形态。比如以考试分数来说,有可能很多学生考得好,而使图形在靠近满分的地方有集中的情况;也可能有很难的题目,使得会做和不会做的学生被区分出来,造成图形有两个尖峰。用眼睛观察图形之后,再描述你看到了什么。

茎叶图

直方图并不是用图形展示分布的惟一选择。数据不很多的时候,画茎叶图比较快,而且呈现更多详细的信息。



眼睛到底看到什么

我们把柱状图和直方图里面的柱体都画成一样宽，因为我们的眼睛会对面积起反应。这样说大致正确。统计学家克利夫兰 (William Cleveland) 的详尽研究显示出，我们的眼睛“看到”的柱体大小，是和它的面积的 0.7 次方成比例。比如说，假设一个象形图里的某个图形是另外一个图形的两倍宽以及两倍高，则较大图形的面积是较小图形的 4 倍。但是我们会把大的图形看成是小的图形的 2.6 倍，因为 4 的 0.7 次方是 2.6。

茎叶图

如何画茎叶图：

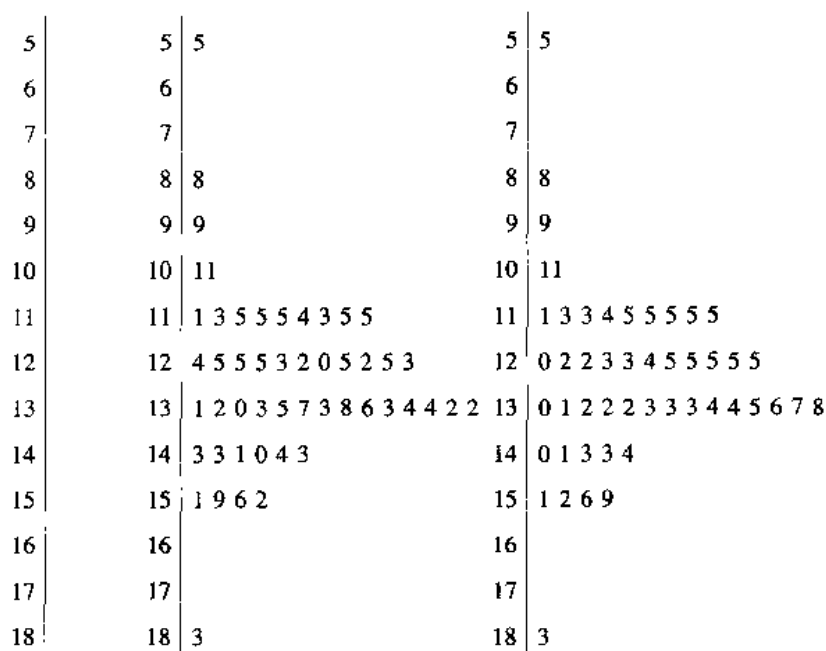
1. 把每个观测值分成茎和叶两个部分，茎包括除了最后一位数字(最右边那个)之外的所有数字，叶就是最后那一位数字。茎视实际需要可以是任何位数，而叶子只能是一位数。
2. 把茎由小到大，从上往下写成一直行，并且在这一直行右边画一条直线。
3. 把每片叶子写在它所属的茎的右边，由小到大排成一列。

例 5 “65 岁以上”资料的茎叶图

对表 11.1 里面的“65 岁以上”百分比来说，观测值的整数部分是茎，最后一位数字(小数第一位)就是叶。阿拉巴马州的数字 13.1 中，13 是茎，1 是叶。茎需要是多少位就是多少位，但是每片叶子只能是一位数。图 11.6 显示了对应表 11.1 的数据画茎叶图的步骤。首先画出发茎。然后就把表里面的数字，逐一加到茎上去做叶子。最后把对应于同一个茎的叶子，由小到大重新排列。

茎叶图其实像是侧躺的直方图。图 11.6 的茎叶图简直就像图 11.3 的直方图，因为直方图所选择的分组，和茎叶图中的茎完全一样。而图 11.2 的茎叶图，其分组组数(即茎数)却是对应同组资料的直方图，即图 11.1 的几乎两倍。解释茎叶图和解释直方图一样，要寻找整体形态，以及有无异常值。

直方图要如何分组由你决定，但茎叶图的组(即茎)却没得选择。你可以把数据四舍五入之后，得到较适合当做叶子的最后一位数，以增加一些弹性空间。数据的位数较多时适合这样处理。比如说，密歇



第一步：写出茎 第二步：加上叶子 第三步：把叶子照顺序排

图 11.6 替表 11.1 的数据画茎叶图。百分比的整数部分是茎，小数第一位是叶

根州大专院校收的学杂费长得像：

15 136 美元 1 940 美元 12 339 美元 6 960 美元…

如果我们把最后一位数当叶子，之前所有位数当茎的话，茎叶图里就会有太多的茎了。要造出图 11.2 的茎叶图，我们先把这些数据全部四舍五入到百元那一位：

151 19 123 70…

这几个数字出现在图 11.2 里 15、1、12 及 7 那几根茎上。

茎叶图的主要优点是呈现了实际的观测值。我们可以从图 11.2 的茎叶图中看出来，密歇根最贵的学校收费是 19 300 美元(四舍五入到百元)，而这点从图 11.1 里就看不出来。茎叶图画起来也比直方图快。茎叶图规定要用头一位或头几位数字当做茎，这等于自动选择了组距，因此有可能画出来的图不能有效描述分布。所以资料数量太庞大时，茎叶图就不适用，因为每个茎都会有太多叶子了。



本章重点摘要

一个变量的**分布**告诉我们该变量有什么值，以及那些值出现的频率。要呈现数量变量的分布可以用**直方图**或**茎叶图**。在观测值的个数不多的时候，我们通常喜欢用茎叶图，而资料数量大时才用直方图。

看一个图的时候，要寻找**整体形态**，以及是否有异于整个形态的**偏差**，比如**异常值**。要描述直方图或茎叶图的整体形态，可以用**形状**、**中心**或**离度**。有些分布有简单的形状，比如说是**对称**或者**偏斜**，但也有些分布太不规则，没法了用一个简单的形状来形容。



第 11 章 习题

11.1 闪电来袭 图 11.7 的资料得自对于科罗拉多州闪电发生状况的研究。图中显示的是，一天之中第一次发生闪电的时间的分布。描述一下这个分布的形状、中心和离度。有没有异常值？

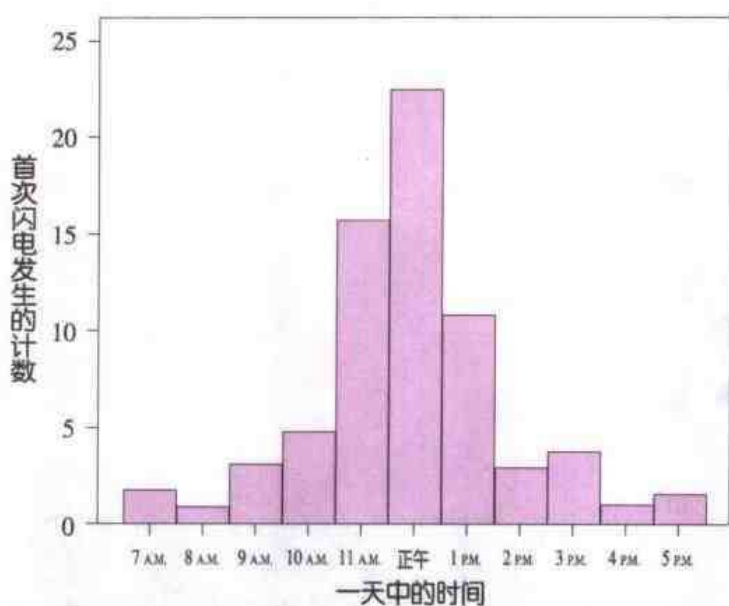


图 11.7 一天之中，第一次闪电发生时间的直方图(资料来自对科罗拉多州的研究)，对照习题 11.1

11.2 年轻人住在哪些州？图 11.8 是美国 50 个州中每一州 25—34 岁居民所占百分比的茎叶图。和图 11.6 的年长居民图一样，茎是百分比的整数部分，而叶是小数第一位。

(a) 也许因为工作机会较少，所以在属洛矶山脉所在的蒙大拿和怀俄明州，年轻居民的百分比最小。这两州的年轻人百分比是多少？

```

10 | 9
11 | 0
12 | 1 3 4 4 6 7 7 8 8 9
13 | 0 0 1 2 4 5 5 5 6 6 7 8 9 9 9
14 | 1 1 2 2 2 3 4 4 4 4 5 7 8 9
15 | 2 4 4 7 8 9 9 9
  
```

图 11.8 美国每一州 25—34 岁居民所占百分比的茎叶图，对照习题 11.2



- (b) 把蒙大拿和怀俄明除外不论, 描述一下分布的形状、中心和离度。
- (c) 年轻人的分布情况, 比起图 11.6 的较年长的居民分布情况, 是不是比较分散?

11.3 主修工程的少数族裔。图 11.9 是 1992—1996 年之间, 在 115 所大学中得到工程博士学位的少数族裔(黑人、西班牙语系、印第安人)学生人数的直方图。简略描述此分布的形状、中心和离度。

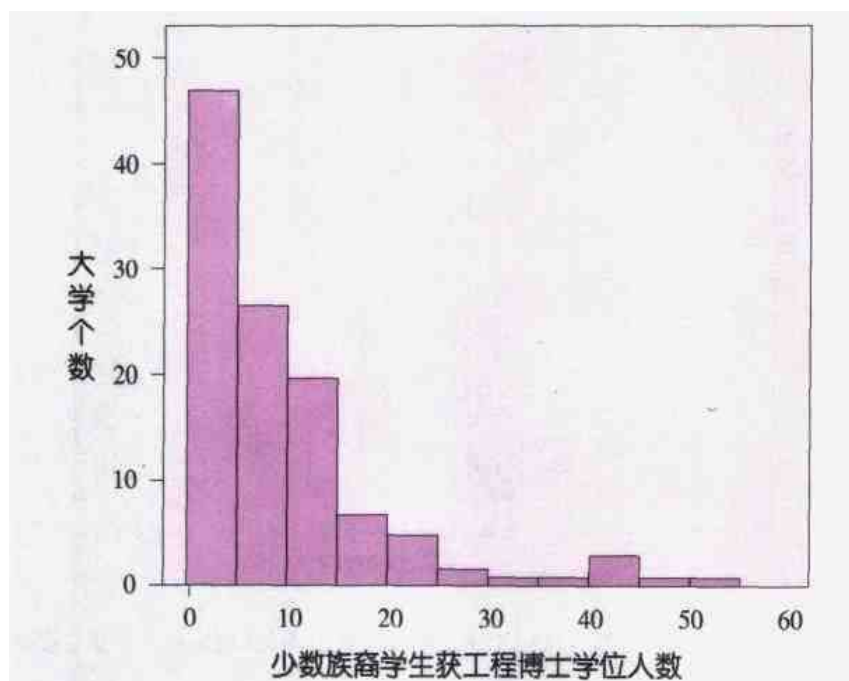


图 11.9 1992—1996 年间, 115 所大学中少数族裔学生获工程博士学位人数的分布, 对照习题 11.3

11.4 股票的获利。一支股票的总获利, 是股价的变化加上股利的配发。通常总获利是用买入时股价的百分比来表示。图 11.10 是一年中当纽约证券交易所中所有挂牌的 1528 支股票总获利分布的直方图。

- (a) 描述一下总获利分布的整体形状。
- (b) 分布大概的中心在哪里? 最小和最大的总获利大致是什么?(这可以用来形容分布的离度。)
- (c) 总获利小于 0, 表示拥有该股票的人亏损。亏损股票大概占多少百分比?

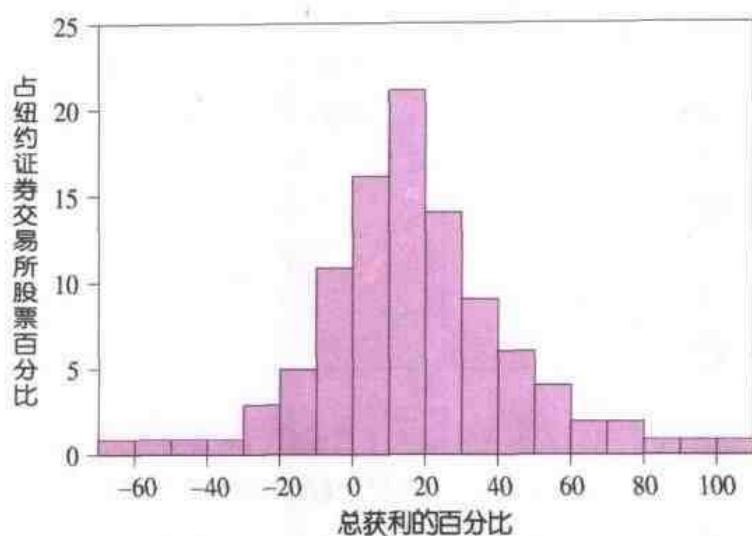


图 11.10 一年之中纽约证券交易所所有股票总获利的分布，对照习题 11.4

11.5 直方图还是茎叶图？说明一下为什么描述 1528 支股票的获利情形时，我们宁愿用直方图而不用茎叶图。

11.6 汽车节油性能比较。美国政府规定汽车制造商必须对每一款车提供城市和公路上的汽油里程数资料。表 11.2 里有 2000 年 32 种中型客车公路里程数(每加仑英里数)的资料。替这些车子的公路里程数画个茎叶图。你认为分布的整体形状如何？中心在哪里(中心是指一半车的里程数比它差，一半比它好的那个值)？有两种车因为公路里程数太低，应该要缴“耗油税”，请问是哪两种？

表 11.2 2000 年中型车的公路里程数

车型	每加仑英里数	车型	每加仑英里数
Acura 3.5RL	24	凌志 GS300	24
奥迪 A6 Quattro	24	凌志 LS400	25
BMW 740i 跑车 M	21	林肯 - 水星 LS	25
别克帝王	29	林肯 - 水星黑貂	28
凯迪拉克凯特拉	24	马自达 626	28
凯迪拉克宝山	28	奔驰 E320	30
雪佛兰 Lumina	30	奔驰 E430	24
克莱斯勒卷云	28	三菱迪蒙特	25
道奇层云	28	三菱 Galant	28
本田雅阁	29	日产顶峰	28



(续表)

车型	每加仑英里数	车型	每加仑英里数
现代索娜塔	28	奥斯摩比北 Intrigue	28
无限 I30	28	绅宝 9-3	26
无限 Q45	23	土星 IS	32
捷豹 Vander Plas	24	丰田佳美	30
捷豹 S/C	21	大众帕萨特	29
捷豹 X200	26	富豪 S70	27

11.7 肥胖流行病。医界权威人士用流行病来形容肥胖在美国的蔓延情形。表 11.3 里列出了参与这项问题研究的 45 州当中,肥胖的成年人所占百分比(NA 代表该州没有资料)。把这个分布用图表示,并简单描述它的形状、中心及离度。

表 11.3 成人当中肥胖者所占百分比(1998 年)

州	百分比	州	百分比	州	百分比
亚拉巴马	20.7	路易斯安那	21.3	俄亥俄	19.5
阿拉斯加	20.7	缅因	17.0	俄克拉何马	18.7
阿里桑纳	12.7	马里兰	19.8	俄勒冈	17.8
阿肯色	NA	马萨诸塞	13.8	宾州	19.0
加州	16.8	密歇根	20.7	罗得岛	NA
科罗拉多	14.0	明尼苏达	15.7	南卡罗来纳	20.2
康涅狄洛	14.7	密西西比	22.0	南达科他	15.4
特拉华	16.6	密苏里	19.8	田纳西	18.5
佛罗里达	17.4	蒙大拿	14.7	德州	19.9
佐治亚	18.7	内布拉斯加	17.5	犹他	15.3
夏威夷	15.3	内华达	NA	佛蒙特	14.4
爱达荷	16.0	新罕布什尔	14.7	弗吉尼亚	18.2
伊利诺伊	17.9	新泽西	15.2	华盛顿	17.6
印第安纳	19.5	新墨西哥	14.7	西弗吉尼亚	22.9
艾奥瓦	19.3	纽约	15.9	威斯康星	17.9
堪萨斯	NA	北卡罗来纳	19.0	怀俄明	NA
肯塔基	19.9	北达科他	18.7		



11.8 洋基队的薪水 表 11.4 列出 1999 年球季开打那天, 纽约洋基棒球队球员的薪水。用这组数据画个直方图。该分布是大致对称、右偏或左偏?

表 11.4 1999 年纽约洋基队员的薪水

威廉斯	9 857 143	柯蒂斯	2 000 000
克莱门斯	8 250 000	赫南德兹	1 850 000
科恩	8 000 000	内尔逊	1 816 667
奥尼尔	6 250 000	苏久	800 000
诺布洛克	6 000 000	门多萨	375 000
佩蒂特	5 950 000	格里姆斯利	370 000
布罗西斯	5 250 000	波萨达	350 000
杰特	5 000 000	诺提	300 000
戴维斯	4 333 000	斯潘塞	204 050
马丁内斯	4 300 000	费格	203 650
里韦拉	4 250 000	勒地	202 850
吉拉弟	3 400 000	贝林杰	200 000
伊拉布	3 125 000	艾那森	200 000
斯坦顿	2 016 667	哲森贝克	200 000

11.9 写作风格的统计 数值资料可以分辨出不同的写作种类, 有时甚至可以分辨出不同作者。以下是学生所搜集来的《大众科学》(Popular Science)杂志里的文章所使用, 长度为 1 到 15 个字母的字所占百分比:

长度	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
百分比	3.6	14.8	18.7	16.0	12.5	8.2	8.1	5.9	4.4	3.6	2.1	0.9	0.6	0.4	0.2

- (a) 为这个分布画直方图。描述一下它的形状、中心和离度。
- (b) 《大众科学》用字长度的分布, 和图 11.5 中莎士比亚剧中用字长度的分布比较起来怎么样? 特别注意一下很短的字(2—4 个字母)以及很长的字(超过 10 个字母)。

11.10 左偏 画一个左偏分布的直方图。假设你和你的朋友把口袋里的零钱通通掏出来, 然后把零钱上的年份记录下来。这些年份的分布会是左偏的。请说明原因。



11.11 会是什么形状?你觉得职业棒球大联盟中 30 个球队的整队球员薪水的分布会是大致对称、明显右偏还是明显左偏?为什么?

11.12 东部各州的西语裔。以下是 1997 年密西西比河以东美国各州,西班牙语系后裔在各州人口中所占百分比:

亚拉巴马	0.1	马里兰	3.5	宾州	2.5
康涅狄格	7.9	马萨诸塞	5.9	罗得岛	6.2
特拉华	3.3	密歇根	2.6	南卡罗来纳	1.2
佛罗里达	14.4	密西西比	0.8	田纳西	1.1
乔治亚	2.8	新罕布什尔	1.4	佛蒙特	0.8
伊利诺伊	9.9	新泽西	11.9	弗吉尼亚	3.5
印第安纳	2.3	纽约	14.2	西弗吉尼亚	0.6
肯塔基	0.8	北卡罗来纳	2.0	威斯康星	2.5
缅因	0.7	俄亥俄	1.5		

替这些数据画一个茎叶图。描述一下分布的整体形态。有没有异常值?

11.13 一根热狗有多少卡路里?《消费者报告》(Consumer Reports)杂志报告了 17 个品牌的肉类热狗每根所含卡路里数:

173	191	182	190	172	147	146	139	175
136	179	153	107	195	135	140	138	

把热狗卡路里含量的分布画成茎叶图,并且大致描述一下分布的形状。大部分品牌的肉类热狗都是用猪肉混合牛肉做成的,也可以有家禽的肉,但依政府规定禽肉不得超过 15%。惟有一个组成成分和别人不一样的品牌是“瘦身小牛肉热狗”,你认为你的茎叶图上面哪一点代表这个品牌?

11.14 变迁中的美国年龄分布。一个国家里面人口的年龄分布,对于该国的经济和社会状况有强烈的影响。表 11.5 显示了 1950 年和 2075 年美国居民的年龄分布,单位是百万人。1950 年的资料来自该年的普查,而 2075 年的资料是普查局所做的预测。



表 11.5 1950 年和 2075 年美国的年龄分布(百万人)

年龄层	1950	2075
不到 10 岁	29.3	34.9
10—19 岁	21.8	35.7
20—29 岁	24.0	36.8
30—39 岁	22.8	38.1
40—49 岁	19.3	37.8
50—59 岁	15.5	37.5
60—69 岁	11.0	34.5
70—79 岁	5.5	27.2
80—89 岁	1.6	18.8
90—99 岁	0.1	7.7
100—109 岁	—	1.7
总数	151.1	310.6

- (a) 因为 2075 年的总人口比 1950 年的大得多, 所以用每一组的百分比来比较, 会比用计数比较清楚得多。造一个每个年龄层占总人口百分比的表, 1950 和 2075 年的资料都要包括在内。
- (b) 画一个 1950 年年龄分布的直方图(用百分比)。描述一下分布的主要特色。特别留意儿童所占百分比。
- (c) 画一个 2075 年预测年龄分布的直方图。横轴和纵轴上用的刻度, 要和(b)中的一样, 以利比较。

在 1950 年和预测的 2075 年这 125 年间, 美国年龄分布的最大变化是什么?

11.15 贝比鲁斯的全垒打。以下是贝比鲁斯于 1920—1934 年身为纽约洋基队球员的 15 年中, 每一年的全垒打数:

54 59 35 41 46 25 47 60 54 46 49 46 41 34 22

画一个茎叶图。这个分布是大致对称、明显偏斜还是两者皆非? 正常情况贝比鲁斯一年大约打出几支全垒打? 他 1927 年著名的 60 个全垒打是不是异常值?

11.16 并列的茎叶图(back to back stemplot). 目前大联盟单季全垒打纪录是由圣路易红雀队的麦奎尔所保有。以下是 1987—1999 年之间, 麦奎尔的全垒打数:



49 32 33 39 22 42 9 9 39 52 58 70 65

一个并列的茎叶图有助于我们比较两个分布。像平常一样把茎的部分写好，但是除了在它右边画一条直线外，左边也要画一条。在右边直线的右边画上鲁斯的叶子(习题 11.15)；在左边直线的左边画上麦奎尔的叶子。每一支茎上的叶子的排列顺序，是以茎为中心，愈往外数字愈大。然后写下你对鲁斯和麦奎尔这两位全垒打王的简短比较。麦奎尔在 1993 年受伤，而 1994 年有职业棒球队员大罢工。这些事件在数据中以什么面貌出现？

11.17 倾盆大雨 1994 年的 7 月 6 日，在佐治亚州阿美里克斯市的土地上下起了 21.1 英寸的雨。那是佐治亚州有史以来，24 小时雨量的最高记录。表 11.6 当中是到 1998 年为止，美国每一州所有气象站曾纪录的 24 小时降水量中的最大值。不同的州之间，得到的纪录有很大差别：飓风会给大西洋沿岸带来大量雨水，而多山的西部基本上很干燥。画一个图来呈现美国各州纪录的分布。大致形容一下这个分布。

表 11.16 美国各州 24 小时降水纪录(英寸)

州	英寸	州	英寸	州	英寸
亚拉巴马	32.52	路易斯安那	22.00	俄亥俄	10.75
阿拉斯加	15.29	缅因	13.32	俄克拉何马	15.68
阿里桑纳	11.40	马里兰	14.75	俄勒冈	11.65
阿肯色	14.06	马萨诸塞	18.15	宾州	34.50
加州	26.12	密歇根	9.78	罗得岛	12.13
科罗拉多	11.08	明尼苏达	10.84	南卡罗来纳	17.00
康涅狄格	12.77	密西西比	15.68	南达科他	8.00
得拉华	8.5	密苏里	18.18	田纳西	11.00
佛罗里达	38.70	蒙大拿	11.50	德州	43.00
佐治亚	21.10	内布拉斯加	13.15	犹他	6.00
夏威夷	38.00	内华达	7.13	佛蒙特	8.77
爱达荷	7.17	新罕布什尔	10.38	弗吉尼亚	27.00
伊利诺伊	16.91	新泽西	14.81	华盛顿	14.26
印第安纳	10.50	新墨西哥	11.28	西弗吉尼亚	19.00
艾奥瓦	16.70	纽约	11.17	威斯康星	11.72
堪萨斯	12.59	北卡罗来纳	22.22	怀俄明	6.06
肯塔基	10.40	北达科他	8.10		

第 12 章

用数字描述分布

多受些教育划算吗?

平均来说,受较多教育的人赚的钱比受较少教育的人多。多多少呢?有一个简单的表示方法,是比较中位收入(median income),这是指有一半人赚得比这个数目多,另一半人赚得比这个数目少。以下是 4 种不同教育程度的成人的中位收入(年薪):

高中毕业	大学毕业	学士学位	更高学位
16 297 美元	18 988 美元	32 581 美元	47 000 美元

这些数字来自 1999 年 3 月当前人口调查(CPS)所访问到的 71 512

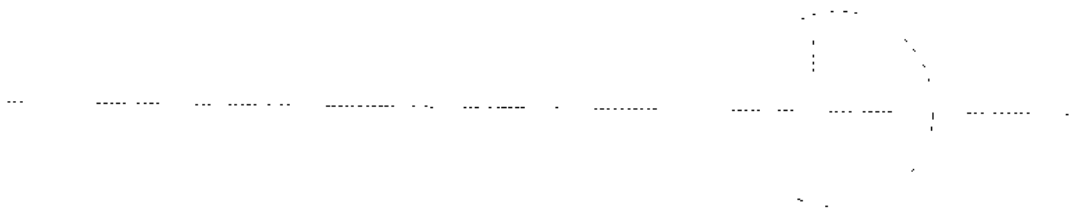


位成人。每年 3 月 CPS 都会对大众的收入做详细调查。把 71 512 个收入浓缩成 4 个数字，可以让我们很快看到教育程度和收入之间的关系。比如说，大学毕业的中等收入者，赚的钱差不多是高中毕业中等收入者的两倍。

我们知道收入的差距很大。事实上，CPS 样本中，高中毕业就没再读书的 31 621 个人里面，年薪最高的是 498 606 美元。中位数 (median) 只比较 4 个收入分布的中心。我们能不能也用很少的数字来描述离度呢？不过，在 31 621 人如此庞大的群组里，当中的最高薪水和最低薪水所能提供的信息有限。不如计算每种教育程度收入分布位于中间那一半的高低差距。以下就是这些资料：

高中毕业	大学毕业	学士学位	更高学位
7 412—29 000 美元	7 803—33 150 美元	16 941—53 061 美元	28 075—75 162 美元

在 71 512 个收入中所含的信息现在更清楚了：读过大学却没拿学位，对于收入的帮助不大，而有学士学位的人赚的钱多很多，有更高学位的人又赚得更多。但是个人之间变异很大——有些很有钱的人从来没上过大学。最后要谨记，这个观测研究不能对因果关系提供任何信息。受高等教育的人通常较聪明，较有野心，且原本家境就较好，所以这些人即使没有学位，仍然可能赚得多。



1998 年的夏天，麦奎尔和索沙激烈角逐美国职业棒球单季全垒打纪录，成为大众关注的焦点。最终由麦奎尔以 70 支全垒打刷新纪录。麦奎尔这项最新成就，相较于他在职业棒球生涯中的全垒打纪录，表现如何？以下是麦奎尔从 1987 年（他的职业棒球生涯第一年）到 1999 年之间的全垒打数：

1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
49	32	33	39	22	42	9	9	39	52	58	70	65



0	9 9
1	
2	2
3	2 3 9 9
4	2 9
5	2 8
6	5
7	0

图 12.1 麦奎尔在
头 13 季职业棒球生涯中
所击出全垒打的茎叶图

图 12.1 的茎叶图展示了这组数据。分布的形状有点不规则，但大致可形容为右偏，单峰而左方有两个异常值。我们可以解释那两个异常值：麦奎尔在 1993 年受伤，而 1994 年发生过球员罢工事件。

一个图再加上几个字，就可以把麦奎尔的全垒打事业描述得很清楚。但是要描述只有高中毕业的 31 621 个人的收入，只用言语描述并不是很恰当。我们需要用数字来代表分布的中心以及离度。

中位数和四分位数

我们在比较不同教育程度者收入的时候，用了很简单而有效的方法来描述中心及离度：也就是中位数及四分位数(quartile)。中位数位于一组数据的正中间，也就是把观测值分隔成数字较小的一半和数字较大的一半的那个值。介于第一四分位数(first quartile)及第三四分位数(third quartile)之间的，就是观测值的中间那一半。四分位数名称的由来，是因为两个四分位数加上中位数，正好把观测值分成四份：四分之一位于第一四分位数之下，二分之一低于中位数，而四分之三低于第三四分位数。这只是基本概念，要实际找到这些数字，还得有个规则来落实这些概念。

例 1 求中位数

要找出麦奎尔在 13 个球季全垒打总数的中位数，得先把这些数字从小到大排列：

9 9 22 32 33 39 **39** 42 49 52 58 65 70

粗体的 39 是最中间的观测值，有 6 个观测值在它左边，6 个在它右边。当观测值的总个数 n 是奇数的时候，从小排到大的数字中，总是有一个是在正中间的，这就是中位数，此例的中位数 $M = 39$ 。

我们也许可以把麦奎尔的纪录和纽约洋基队外野手马里斯(Roger Maris)的比一比，麦奎尔所破的单季纪录，原先就是由马里斯所保持。以下是马里斯在美国联盟



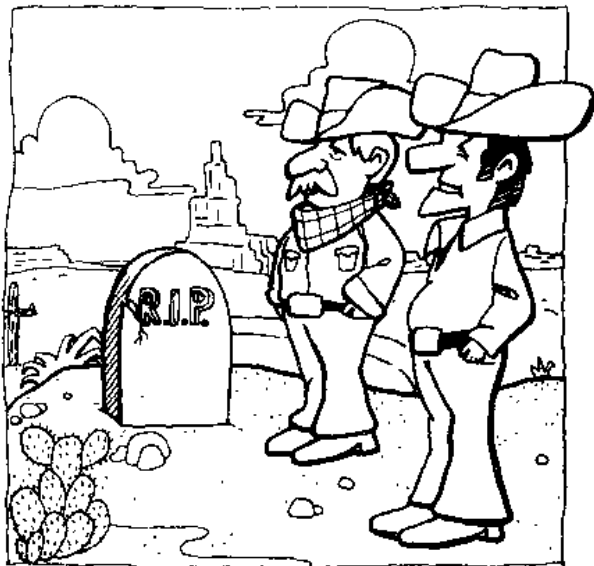
十年当中的全垒打计数，从小排到大：

8 13 14 16 **23** **26** 28 33 39 61

当 n 是偶数的时候，没有一个在正中间的观测值，但有一对在正中间的观测值：粗体的 23 和 26，他们各有 4 个观测值在其外侧。中位数就取在中间这一对的中点。所以马里斯的中位数是：

$$M = \frac{23 + 26}{2} = \frac{49}{2} = 24.5$$

数字照顺序排好后，有一个方法可以很快找到中位数的所在：从头算起一直到 $(n+1)/2$ 的位置。你可以试试看。对麦奎尔来说， $n=13$ ，而 $(n+1)/2=7$ ，所以中位数位于从头数起第 7 个位置。对马里斯来说， $n=10$ 而 $(n+1)/2=5.5$ ，这代表“位于第 5 和第 6 个数的中间”，所以 M 就是这两个数字的平均。而这个“ $(n+1)/2$ 规则”在有很多观测值的时候尤其好用。 $n=31\,621$ 个收入的中位数，是排序之后的第 15 811 个数。不过要注意 $(n+1)/2$ 并不等于中位数 M ，而是指在观测值排序后，中位数所在的位置。



“没错，老包会溺毙完全是因为不懂统计，他还以为只要知道河的平均深度就成了呢。”

注 R. I. P 是 Rest in Peace 的简写，愿他安息之意

• 中位数 M

中位数 M (median, M) 是一个分布的中间点，也就是一半观测值比它小，一半比它大的那个数。要找分布的中位数，步骤如下：

1. 把所有观测值排顺序，由小到大。
2. 若观测值个数 n 为奇数，中位数 M 就是排序后观测值最中间的一个。要找中位数的位置，只要从头数起，数到第 $(n+1)/2$ 个位置即可。
3. 若观测值个数 n 为偶数，中位数 M 就是排序后最中间的两个观测值的平均。要找中位数的位置，仍然是从头数到第 $(n+1)/2$ 个位置即可。



当观测值很多的时候,你可以利用 $(n+1)/2$ 的规则来找到四分位数的位置。因为31 621个收入的中位数,是从小到大排顺序之后的第15 811个数字,所以第一四分位数是中位数之前的15 810个收入的中位数。我们用 $(n+1)/2$ 规则来找出第一四分位数的位置, $n=15\,810$:

$$\frac{n+1}{2} = \frac{15\,810+1}{2} = 7905.5$$

收入排序后的第7 905个数字和第7 906个数字的平均是7 412美元,这个数字在本章开头的例子中就出现过了。

五数综合及箱形图

最小和最大的观测值对于整体分布可以提供的信息很有限,但是它们提供了关于分布尾部的信息,而当我们只知道中位数和四分位数时,是对分布的尾部毫无所知的。要迅速掌握分布的中心和离度的话,可以把这五个数字综合起来。

五数综合

一个分布的**五数综合**(five-number summary),从小写到大,包括:最小数、第一四分位数、中位数、第三四分位数及最大数。用符号表示的话,五数综合是:

$$\text{最小数} \quad Q_1 \quad M \quad Q_3 \quad \text{最大数}$$

这五个数字对于分布的中心和离度,提供了大致完整的描述。麦奎尔全垒打计数的五数综合是:

$$9 \quad 27 \quad 39 \quad 55 \quad 70$$

而马里斯的是:

$$8 \quad 14 \quad 24.5 \quad 33 \quad 61$$

根据一个分布的五数综合可以画出新的图,也就是箱形图(boxplot)。图12.2中显示出两组全垒打数据的箱形图比较。



• 箱形图

箱形图(boxplot)是显示出五数综合出的图。

- 箱形图中间的箱体,是从第一四分位数延伸到第三四分位数。
- 箱体里的直线标示出中位数的位置。
- 箱体两头有直线往外延伸到最小数和最大数。

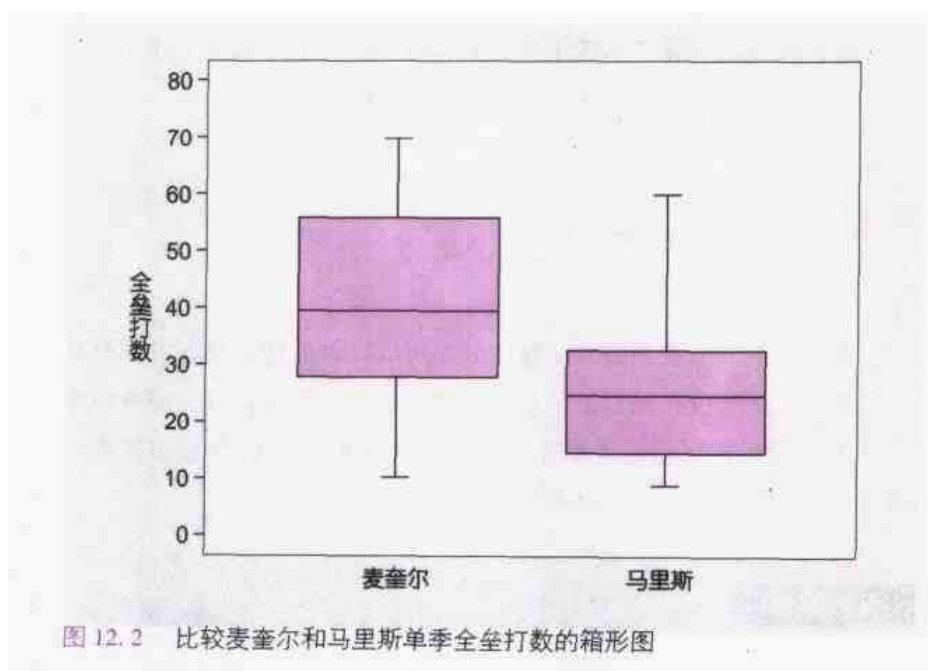


图 12.2 比较麦奎尔和马里斯单季全垒打数的箱形图

箱形图可以直着画也可以横着画,但要记得在图中标示出数字刻度。检视箱形图的时候,要先找出中位数的位置,这就是分布的中心所在。然后看看离度,两个四分位数的距离,显示出中间一半数据的分散状况,而箱形图的两端(最小数和最大数)则显示出整组数据的分散情况。从图 12.2 可以看出,如果以中位数和箱形图中箱体(涵盖中间一半的数据)的位置来代表一般表现的话,麦奎尔是比马里斯要出色的。

因为箱形图包含的细节比直方图及茎叶图少,所以它的最佳用途是用来同时比较至少二个分布,就像图 12.2 那样。不过对于这样少的观测值来说,画一个并列的茎叶图会更好(参考习题 11.16)。因为从茎叶图可以清楚看出,马里斯在 1961 年的 61 支全垒打的纪录,在他整个生涯中只不过是异常值而已,而这点从箱形图看不出来。让我们来看一个箱形图真正有用的例子。



例3 教育及收入

为了了解不同教育程度的人收入的差别有多大，我们查阅了当前人口调查中每年对收入所做的抽样调查结果。图 12.3 比较了根据 71 512 个观测值所得到的四种不同教育程度的收入分布。这个图的概念和箱形图一样，只有稍做改变。几万个人当中的最高收入数字一定很大。比如说，高中毕业的人当中收入最高的，是 498 606 美元，而在这组教育程度的人里面，只有 5% 的人收入超过 54 481 美元。图 12.3 用分布当中 5% 和 95% 的点，各代替了最低和最高收入。所以只读完高中的那组，箱体上面的直线只延伸到 54 481 为止。

图 12.3 为我们提供了清楚又简单的视觉比较。我们可以看到，对于拥有学士学位和更高学位的人，中位数和中间一半的收入，位置都比较高。而每一组的最低 5% 收入都很低，因为每一组都会有些人，可能因为生病或身心障碍而没有收入，甚至有负的收入。95% 的点，也就是隔开最高 5% 收入的人，在高学历的人当中飙升得超高，这些人包括医师、律师和有工商管理硕士学位的人。

图 12.3 也可以用来说明，为什么通常可以从箱形图看出一个分布是对称还是有偏的。一个对称分布的第一和第三四分位数，距离中位数的距离都是一样远的。而对于大部分的右偏分布来说，第三四分位数距中位数的距离，会超过第一四分位数距中位数的距离。而两头极端的值也有同样的情况，即使顶尖收入的 5% 没有包括在图里面，我们还是可以看到高学位的人的收入分布，有很强的右偏现象。

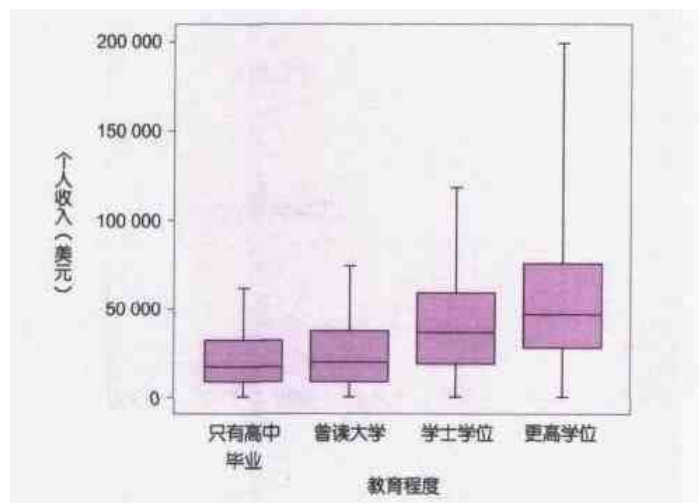


图 12.3 比较不同教育程度成人收入分布的箱形图。每一个箱形图的两端，是分布的 5% 和 95% 的点



统计学上的争议

贫富差距

在繁荣的 20 世纪 80 和 20 世纪 90 年代,美国住户的收入虽然增加,但是贫富差距却也加大了。图 12.4 和 12.5 对于日益增加的差距提供了两种不同的观点。图 12.4 是住户收入的线图,单位是美元,但是经过调整,使得 1 美元在每一年的购买力都一样。三条线分别代表中位收入,第 20 百分位数(percentile)及第 80 百分位数,后二者分隔出了收入最低的五分之一户和收入最高的五分之一户。第 80 百分位数(在 1967—1998 年间增加了 41%)和中位数及第 20 百

分位数之间的差距愈拉愈远,因为后二者都只增加了 20%。

图 12.5 画的是收入最高的五分之一户以及最低五分之一户分别占总收入的多少百分比。收入最低的五分之一户,收入所占的百分比缓慢下降,到 1998 年只占全部收入的 3.6%。收入最高的五分之一户,总收入却增加到占全部收入的 49.2%。收入最高的 5% 住户,收入增加更快,已超过全国住户总收入的 21%。贫富差距情形在美国比其他发达国家要严重,而且差

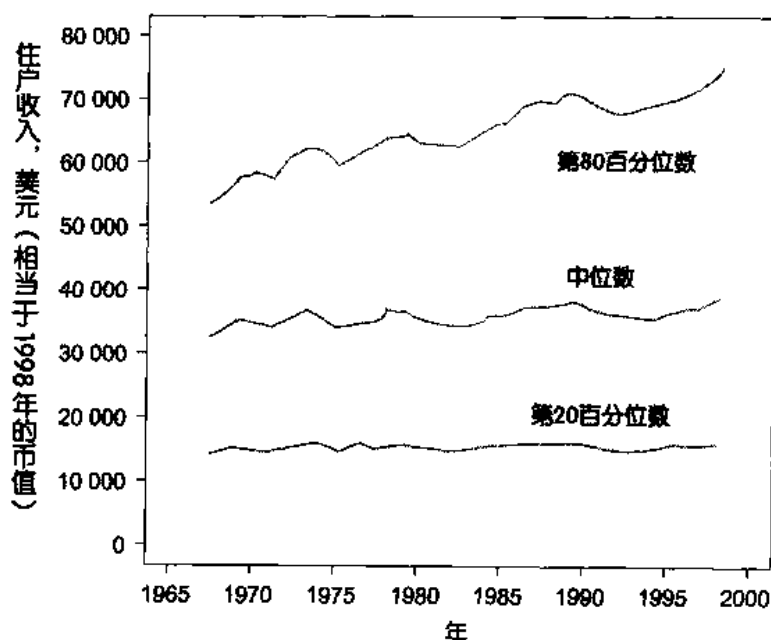


图 12.4 美国住户收入分布中的三个点随时间改变的状况。有 80% 的住户,收入在第 80 百分位数之下,一半在中位数之下,20% 在第 20 百分位数之下。1998 年时,第 20 百分位数为 16 116 美元,中位收入是 38 885 美元,而第 80 百分位数是 75 000 美元



距还在加大当中。

这是很复杂的议题，包含许多互相矛盾的数据以及一些隐藏的事项。左派人士想要减少贫富差距，而右派人士认为有钱人的高收入是他们应得的。我必须指出，这里牵涉到重要的统计观点的问题。图12.4和12.5中报告的是“横截面”资料(cross section data)，也就是每一年的所有

住户资料。如果追踪各住户随着时间改变的“历时”资料(longitudinal data)，可能会看到不一样的情况。我们可以观察买玛和唐雅这对年轻夫妇，他们先是半工半读，然后贷款去读研究生。那时他们的收入属于最低的五分之一。在他们毕业之后收入急剧增加，到40岁的时候，他们已跻身于最高五分之一的高收入族。许多低收入户

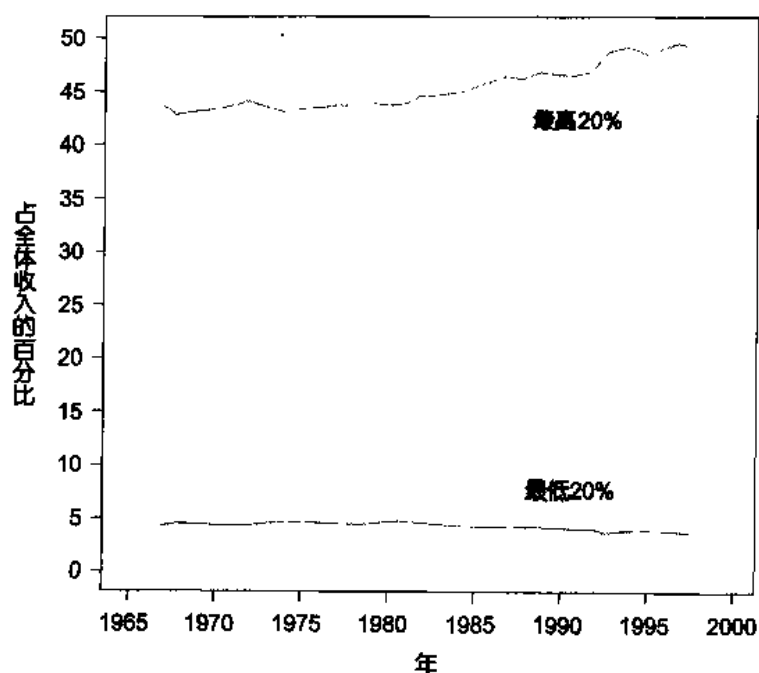


图12.5 最高收入20%及最低收入20%的住户的收入，占全体住户收入之百分比随时间变化的状况。1998年时，顶尖20%收入的住户的收入，将近占全体收入的一半

只是暂时性的低收入而已。

历时研究很费钱，因为必须追踪同样的住户许多年。而且因为有些住户会半途退出，所以研究结果容易有偏差。有一项对所得税申报书做的研究发现，收入最低的五分之一住户当中，十年之后只有14%仍然属于最低收入的五分之一。不过，真正最穷的人根本不申报所得税。另一项研究是以5岁以下儿童为对象。分别从1971

年及1981年开始的研究显示，原本在底部五分之一低收入户的儿童。有60%在十年后仍然属于底部五分之一的低收入户。的确有许多人在年龄渐增后由穷转富，但是也仍然有很多住户一直都很穷。而不幸的是，有许多儿童生在这些穷人家。图12.6显示出贫富差距加大的另一方面：生活在贫穷家庭中的儿童增加了。

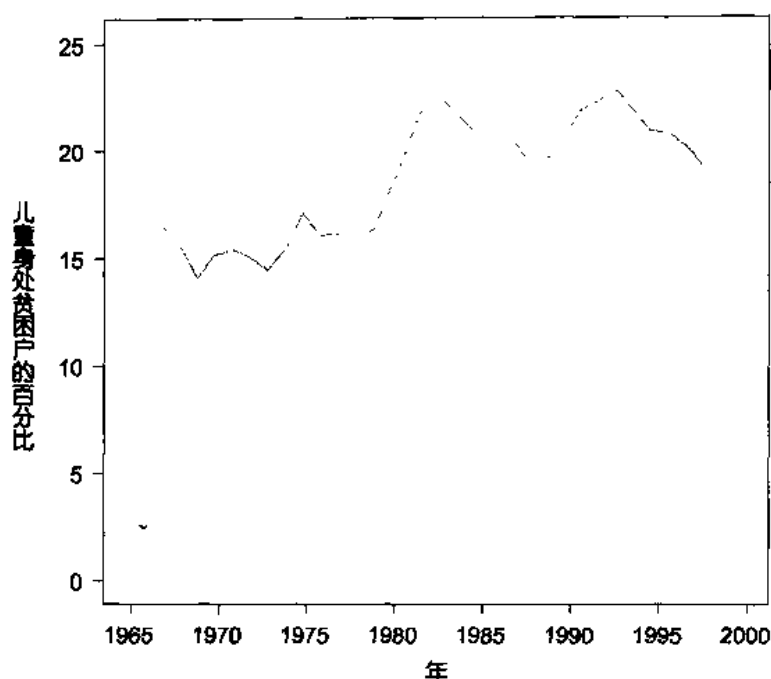


图 12.6 居住于政府认定为贫困户的儿童百分比, 随时间变化的情况。1998 年时, 美国有 18.9% 的儿童身处贫困户

平均数和标准差

五数综合并不是描述分布最常用的数值。最常用的是拿平均数 (mean) 来度量中心, 加上用标准差 (standard deviation) 来度量离度。平均数我们很熟悉——它就是一般的观测值平均而已。标准差的概念, 是找出观测值距平均数的平均距离。但是标准差所代表的“平均距离”, 却不是用一个简单明了的公式计算的。我们会把公式列出来, 但是你可以就把标准差想成是“与平均数相距的平均距离”, 而把计算的部分留给计算机去做。

实际要算的时候, 你可以把数据输入你的计算机, 然后按平均数键及标准差键。你也可以用现成的软件来计算 \bar{x} 及 s 。通常要算距离平方的平均, 都是除以 $n-1$ 而不是 n , 这背后有很好但是具专门性的理由。有很多计算机有两个标准差键, 你可以选择要除以 n 还是除以 $n-1$, 记得要选 $n-1$ 。



平均数和标准差

一组观测值的平均数(mean) \bar{x} (读成 X-bar), 就是该组观测值的平均。要找出 n 个观测值的平均数, 只要把那组值全部加起来再除以 n 即可:

$$\bar{x} = \frac{n \text{ 个观测值的和}}{n}$$

标准差 s (standard deviation s) 度量的是观测值与平均数间的平均距离。计算的方法是先算出各距离平方后的平均值, 再取平方根。要算出 n 个观测值的标准差, 步骤如下:

1. 先找出每个观测值距平均数的距离, 并把这个距离平方。
2. 把所有的距离平方加起来, 并除以 $n-1$ 。所得到的距离平方的“平均”, 叫做方差 (variance)。
3. 标准差 s , 是再把这个方差取平方根。

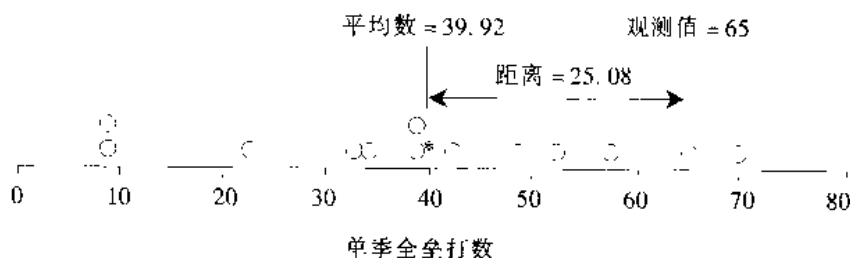


图 12.7 麦奎尔的全垒打数和其平均数(*)以及其中一个观测值距平均数的距离, 把标准差想成是这些距离的平均

例 4 求平均数和标准差

麦奎尔在他人联盟棒球生涯的头十三季的全垒打数如下:

49 32 33 39 22 42 9 9 39 52 58 70 65

这些观测值的平均数是:

$$\begin{aligned}\bar{x} &= \frac{n \text{ 个观测值的和}}{n} \\ &= \frac{49 + 32 + \cdots + 65}{13}\end{aligned}$$



$$= \frac{519}{13} = 39.92$$

图 12.7 把数据用数线(number line)上的点来表示, 而整组数据的平均数, 用星号(*)标示出来。双向箭头显示出其中一个观测值距平均数的距离。标准差 s 的概念, 是要把全部 13 个距离平均起来。如果要手算标准差, 可以用下面这个表的表示方法:

观测值	距平均数距离之平方
49	$(49 - 39.92)^2 = (9.08)^2 = 82.45$
32	$(32 - 39.92)^2 = (-7.92)^2 = 62.73$
65	$(65 - 39.92)^2 = (25.08)^2 = 629.01$
...	... 和 = 4 438.97

求平均得到:

$$\frac{4\,438.97}{12} = 369.91$$

请注意我们求“平均”的时候, 除数比观测值的个数要少 1。最后, 标准差就是这个数字的平方根:

$$s = \sqrt{369.91} = 19.23$$

比这些计算细节更重要的, 是能显示出为什么标准差可以度量离度的性质。

• 标准差 s 的性质

- s 度量的是以 \bar{x} 为中心的离度。只有在你用 \bar{x} 来描述分布中心时, 才可以用 s 来描述离度。
- 只有在没有离度的时候, s 才会等于 0。而这种情况只发生在所有观测值都相同的时候。所以标准差为零代表观测值完全没有散布(全都在同一点), 否则 s 必然大于零。当观测值离平均数散布得愈远时, s 就愈大。



例5 投资入门

收入的例子已经够多了。现在举个例子来讨论你把收入赚进来后,应该拿它怎么办。投资的首要法则中有一条是,肯冒多一点险、获利就比较高,至少长期下来平均是如此。金融界的人度量风险,是看一项投资的获利,其不可预测的程度如何而定。钱存在有政府保险的银行,而且用固定利率,一点儿风险也没有,因为获利多少是完全确定的。一家新公司的股票则可能一周之内暴涨,下一周又暴跌。它的风险很高,因为你没法预测当你一旦要卖时,它将会值多少钱。

投资者应该利用统计思考。你可以以一项投资年获利的分布来做评估。这代表必须了解获利形态的中心以及离度。只有不成熟的投资者,才会只关心平均获利高不高,却不管风险如何,也就是不管获利的散布广不广,变化大不大。金融专家用平均数和标准差来描述投资的获利状况,长久以来他们都觉得标准差太复杂,不适合向一般人众提及,不过现在你会开始注意到,在共同基金的定期报告中,都曾提出标准差是多少。

为了说明,我们来看看以下所列出的,1950—1999年的50个年头,三种投资年获利的平均数和标准差:

投资	平均获利	标准差
短期国库券	5.34%	2.96%
长期国库券	6.12%	10.73%
股票	14.62%	16.32%

你可以看到,平均获利上升,风险(变异)就跟着上升,这恰和金融理论所说的相印证。短期和长期国库券都是借钱给美国政府的方法。短期国库券一年就偿还,其获利每年随利率变化而不同;长期国库券是30年后偿还,风险比较大,是因为如果利率上升,你拥有的国库券价值就下降。股票的风险更大,它们的获利较高(以长期下来平均而言),但是其间会有许许多多的大起大落。从图12.8的茎叶图就可看出,在资料涵盖的50年当中,股票的年获利有高达50%的,也有低到亏损26%的。

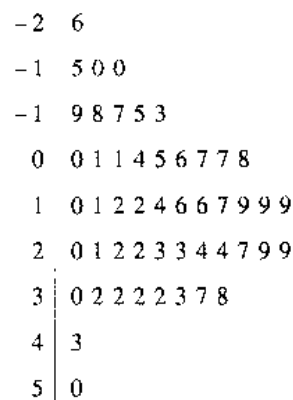


图 12.8 1950~1999 年共 50 年，股票年获利的茎叶图。获利经四舍五入到百分比的个位数。茎的单位是十个百分点，叶子是一个百分点

选择数值描述

五数综合很容易懂，对于大部分的分布而言，它也是最佳的精简描述。平均数和标准差比较难懂，但却比较常用。我们应该怎样决定要用哪一种来描述中心和离度呢？让我们先比较一下平均数和中位数。不论用“中间点”还是“算术平均”(arithmetic average)来描述一组数据的中心，都是很合理的概念，但是二者的概念不同，用处也不一样。最重要的差别是，平均数(算术平均)会因少数极端值而受很大的影响，而中位数(中间点)则不会如此。

例 6 平均数和中位数的差别

表 12.1 是 2000 年洛杉矶湖人篮球队 14 位队员的大概薪水(百万美元)。你可以算一算，平均数是 $\bar{x} = 410$ 万而中位数是 $M = 260$ 万，难怪职业篮球队员都住豪宅。

表 12.1 洛杉矶湖人队 2000 年薪水(百万美元)

球员	薪水	球员	薪水
奥尼尔	17.1	哈珀	2.1
布莱恩特	11.8	格林	2.0
霍利	5.0	乔治	1.0
莱斯	4.5	肖	1.0
费希尔	4.3	萨利	0.8
福克斯	4.2	卢	0.7
奈特	3.1	塞莱斯坦德	0.3

平均数为什么比中位数高这么多呢？图 12.9 是这些薪水的茎叶图，茎是百万美元，可以看出分布是右偏的，且有两个很大的异常值。科比·布莱恩特和沙奎尔·奥尼尔的超高薪水把薪水总和拉高，因此把平均数也拉高了。如



果不计入这两个异常值,其他12位球员的平均薪水只有240万美元而已,但中位数的差别并不会这么大,它只从260万降到205万而已。

只要增加沙奎尔一个人的薪水,就可以把平均薪水加到任何我们想要的数字,因为只要一个异常值一直往上移,平均数就会跟着往上移。但是对中位数来说,沙奎尔的薪水就只不过是分布在最高端的一个观测值而已,把这个数字从1710万改成17100万,并不会对中位数产生丝毫影响。

对称分布的平均数和中位数很接近。事实上,当分布完全对称的时候, \bar{x} 和 M 根本就相等。然而在偏斜分布里,平均数就会离中位数而去,靠向较长的尾部。很多和钱有关的分布,例如收入、房价、财富等,都有很强的右偏现象。平均数可能比中位数大很多。比如说,在本章开始的时候我们谈到过,当前人口调查样本里,高学历人士的中位收入是47000美元。该分布是右偏的,而右边的长尾巴就把平均数给拉高到65220美元了。因为有关钱的数据常常有少数特别大的观测值,所以要描述这类分布常用中位数而不用平均数。

在平均数和中位数之间做选择的时候,要考虑的不只是分布是对称还是偏斜而已。米德尔敦房屋售价的分布无疑会是右偏的,但是如果该市的市议会为了决定税率,而要估计所有房屋的总市值的时候,则对他们有帮助的数字是平均数而非中位数,因为总市值只不过是房屋总数乘上平均数而已,和中位数并没什么关系。

标准差被异常值或偏斜分布的长尾巴拉走的情况,比平均数还要严重。湖人队全体14位队员薪水的标准差是 $s=476$ 万美元,而如果不计入两个异常值,则 s 只等于172万美元。不过四分位数对于少数极端值就不这么敏感。还有一个理由让我们应该避免用标准差来描述偏斜分布:因为一个明显偏斜的分布的两边,散布情形并不一样,所以若只用一个数字,比如像 s ,没有办法恰当地描述离度。而五数综合里有两个四分位数以及最大及最小数,所以比较理想。在大部分情况下,只有在分布大致对称的时候才用 \bar{x} 和 s ,这是

纽约是穷州?

纽约是不是个富州?

纽约州的个人平均收入在美国全部50个州中位居第四,和它的富邻居康涅狄格及新泽西州一起名列前茅(后两州分列一、二名)。但是康涅狄格和新泽西州的住户中位收入分居全国第七和第二名,纽约州却排第二十九,比全国平均的中位收入低许多。这是怎么回事?这只不过是平均数不同于中位数的另一个例子。纽约州有许多收入非常高的居民,把平均收入提高许多。但是它的贫困户比例比新泽西和康涅狄格都要高,使得住户中位收入偏低。纽约州并不有钱——它只是同时拥有非常有钱和非常贫穷的居民这两种极端的例子。



0 | 3 7 8
 1 | 0 0
 2 | 0 1
 3 | 1
 4 | 2 3 5
 5 | 0
 6 |
 7 |
 8 |
 9 |
 10 |
 11 | 8
 12 |
 13 |
 14 |
 15 |
 16 |
 17 | 1

图 12.9 洛
 杉矶湖人队队
 员薪水之茎叶
 图, 数据来自
 表 12.1

较明智的。

• 选择适当的综合数值描述

平均数和标准差会受异常值或偏斜分布的长尾巴严重影响, 而中位数和四分位数则比较不受影响。

要描述偏斜分布, 或者有异常值的分布, 五数综合通常要比平均数和标准差更合适。只有在分布大致对称又没有异常值的时候, 才用 \bar{x} 和 s 。

那我们干嘛还要花费精力在标准差上面呢? 有一个答案会在下一章出现: 对于一种叫做正态分布(normal distributions)的重要对称分布来说, 平均数和标准差是中心和离度的理所当然的量度。

请记住, 图形可以对分布提供最清楚的整体情况。中心和离度的数值量度告诉我们分布的某些特征, 但是并没有描述整个分布的形状。比如说, 数值综合就没告诉我们, 分布是不是有好几个高峰(peak), 或者中间有没有空档。因此切记: 每次拿到资料都应该先画图。

网络寻奇

本章中的许多例子都和收入有关。你如果访问美国普查局的网站: www.census.gov, 就可以找到美国最新的收入资料及相关议题。要找 *Money Income in the United States* (在首页上点击“income”)以及 *Poverty in the United States* (点击“Poverty”)。你可以看到这些每年发表的资料的完整内容及许多的表, 还可以看到许多实际资料, 比如图 12.3 里 71 512 位成人的收入和教育程度。

你也可以上《统计学的世界》原文版所设的网站: www.whfreeman.com/sec, 进到“统计应用小程序”(Statistical Applets), 再选择“平均数和中位数”小程序(“Mean and Median” applet), 可以实际比较平均数和中位数的变化情形。你可以按鼠标左键把数据点进去, 然后用鼠标把一个异常值往上拖, 看着平均数跟在后面追。



本章重点摘要

要描述一组数据，一定要先画图，然后才加进经过谨慎选择，可以把该组数据的特性综合出来的数字。如果我们的数据只属于单一的数量变量，可以先用直方图或茎叶图来呈现其分布，然后再用一些数字来描述该分布的**中心及离度**。

描述中心和离度有两种常用方式：**五数综合**以及**平均数和标准差**。五数综合里包含了用来度量中心的中位数，以及两个用来描述离度的**四分位数**加上最小和最大观测值。中位数位于所有观测值的中间位置。**平均数**是所有观测值的平均、**标准差**度量离度，它差不多是距平均数的平均距离，所以用标准差的时候，一定是用平均数来度量中心。

平均数和标准差都会因为少数异常值而受很大的影响。对于对称分布来说，平均数和中位数差不多一样，但是对偏斜分布来说，平均数就会更加朝着长尾方向移动。总括来说，大部分的分布都适合用五数综合来描述，但是平均数和标准差就只适合用在大致对称的分布上。



第12章 习题

12.1 中位收入 你已读到了美国住户在1998年的中位收入是38 885美元。请用日常用语解释什么叫做“中位收入”。

12.2 平均是多少?美国普查局的出版物《美国的金钱收入》(*Money Income in the United States*)中提供了好几种的“平均收入”。1998年美国住户的中位收入是38 885美元。住户平均收入是51 855美元。而家庭中位收入是46 737美元,家庭平均收入是59 589美元。普查局说明,“住户”成员是所有同住一个住宅单位的人,而“家庭”指的是两个以上住在一起,而具有经由血缘、婚姻或收养而产生关系的人。详细解释为何平均收入高于中位收入,还有为何家庭收入高于住户收入。

12.3 富有的杂志读者。在1999年7月5日,商业杂志《福布斯》报道说,它的读者的住户中位财产是956 000美元。

- (a) 你认为这些住户的平均财富会比956 000美元高还是低?为什么?
- (b) 数据是根据电话联络《福布斯》读者所得样本而来的。我怀疑956 000美元可能比实际中位财富高了些。你想为什么?

12.4 大学学费。图11.2是密歇根州81所大专院校所收学杂费之茎叶图。茎是千元,叶是百元。举例来说,最高的学杂费是19 300美元,在茎叶图上出现在茎19及叶3的位置。

- (a) 替这组密西根州大学学费资料找出五数综合。请注意茎叶图已经把数据按大小排顺序了。
- (b) 平均学费会明显小于中位数、和中位数差不多还是明显大于中位数?为什么?

12.5 年轻人住哪?图11.8是美国50个州中每一州25—34岁居民所占百分比之茎叶图。茎是一个百分点,叶子是十分之一个百分点。

- (a) 从分布的形状来看,平均数和中位数不会差很多。为什么?
- (b) 找出这组数据的平均数和中位数,来证实二者很接近。



12.6 汽油里程数 表 11.2 中有 2000 年中型车的每加仑公路汽油里程(英里)数。

- (a) 假如你没做习题 11.6 的话, 现在画一个这些数据的茎叶图
- (b) 找出公路里程数的五数综合。哪些车的汽油里程数属于最低的四分之一?
- (c) 从茎叶图可以看出关于分布整体形态的一个事实, 是五数综合没法描述的。这是指哪一个事实?

12.7 洋基队薪水 表 11.4 里是纽约洋基棒球队的薪水。你预期这个分布会有怎样的形状? 你认为平均薪水会接近中位薪水、明显高于中位薪水还是明显低于中位薪水? 画一个图并计算平均数和中位数, 来印证你的选择。

12.8 最有钱的 1% 美国个人收入的分布, 右偏的状况非常明显。1997 年美国最高收入 1% 的人, 平均收入和中位收入 ~ 是 330 000 美元, ~ 是 675 000 美元。这两个数字哪个是平均数, 哪个是中位数? 解释你的理由。

12.9 一根热狗有多少卡路里? 美国《消费者报告》杂志报道了以下 17 个品牌单根热狗的卡路里含量:

173	191	182	190	172	147	146	139	175
136	179	153	107	195	135	140	138	

画一个茎叶图并找出五数综合。从茎叶图可以看出一些有关分布的重要信息, 是从综合数值中看不出来的, 请问是哪些信息?

12.10 股票获利 例 5 告诉我们, 财务理论中是用平均数和标准差来描述投资的获利状况。图 11.10 是一年当中纽约证券交易所所有股票获利资料的直方图。若要简单描述这个分布, 用平均数和标准差合适不合适? 为什么?

12.11 工科的少数族裔学生。图 11.9 中, 是 1992—1996 年期间, 美国的 115 所大学中得到工程博士学位的少数族裔学生(黑人、西班牙语系、印第安人)人数的直方图。



- (a) 把全部 115 个观测值从小排到大之后, 五数综合的 5 个数字, 在这 115 个数字中的位置各在哪里?
- (b) 即使没有实际的 115 个数据, 你也可以利用你对(a)小题的解答以及直方图, 来找出近似的五数综合, 请试着做做看。大学若要列名在最高四分之一的話, 大约要授予工程博士学位给几个少数民族裔学生?

12.12 写作风格的统计。以下是《大众科学》杂志的文章中, 用字长度在 1 到 15 个字母分别所占百分比的资料。习题 11.9 曾要求你画出直方图。

长度	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
百分比	3.6	14.8	18.7	16.0	12.5	8.2	8.1	5.9	4.4	3.6	2.1	0.9	0.6	0.4	0.2

稍微用点脑筋, 你就可以从这个表找出字长分布的五数综合。现在就找找看。

12.13 东部各州的移民人数。以下是 1997 年在密西西比河以东美国各州定居的移民人数(以千人为单位):

以这个分布画一个图。描述一下分布的整体形态和异常值。然后选择适当的数值综合并计算出它们的值。

亚拉巴马	1.6	康涅狄格	9.5	特拉华	1.4
佛罗里达	82.3	佐治亚	12.6	伊利诺伊	38.1
印第安纳	3.9	肯塔基	1.9	缅因	1.0
马里兰	19.0	马萨诸塞	17.3	密歇根	14.7
密西西比	1.1	新罕布什尔	1.5	新泽西	41.2
纽约	123.7	北卡罗来纳	5.9	俄亥俄	8.2
宾州	14.6	罗得岛	2.5	南卡罗来纳	2.4
田纳西	4.4	佛蒙特	0.6	弗吉尼亚	19.3
西弗吉尼亚	0.6	威斯康星	3.2		

12.14 东部各州的移民人数。在前一题中, 纽约是分布的较大异常值。请算出上一题数据的平均数和中位数, 一次包括纽约在内, 一次不包括。我们把异常值除外之后, 平均数和中位数何者变化较大?



12.15 各州学生的 SAT 分数。图 12.10 是美国 50 州和哥伦比亚特区的学生在 SAT 测验中，数学平均分数的直方图。从这个分布的特殊形状可以得知，只提出一个中心的量度，比如说平均数或者中位数，对于描述分布来说实在没多大用处。说明一下为什么如此。

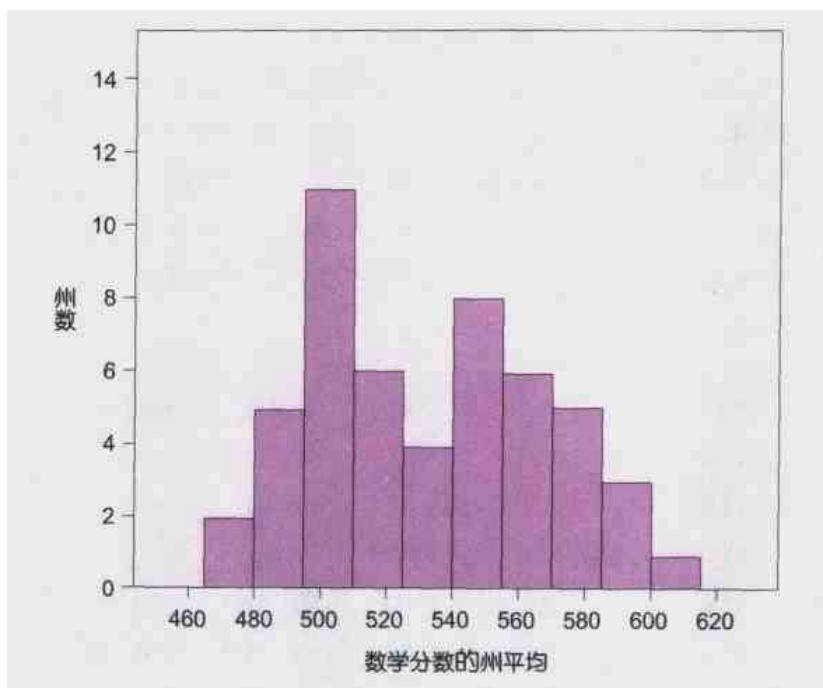


图 12.10 美国 50 州以及哥伦比亚特区学生 1998 年 SAT 数学平均分数的直方图，对照习题 12.15

12.16 高薪运动员。有一项新闻报道说，1998 年 2 月在美国国家篮球协会(NBA, National Basketball Association)名单上的 411 位球员中，只有 139 位“薪水超过联盟的平均薪水”236 万美元。236 万美元是 NBA 球员的平均薪水还是中位薪水？你怎样知道答案的？

12.17 平均数还是中位数？在以下状况中，应该用平均数还是中位数来当做中心的量度？为什么？

- (a) 米德尔敦考虑要对居民征收所得税。市政府想要知道一般市民的收入是多少，以便估计总税基(total tax base)。
- (b) 一位社会学家为了研究米德尔敦典型家庭的生活水平，而估计该城的一般家庭收入。



12.18 平均数还是中位数?你在策划一次聚会, 想要知道必须买多少罐汽水。假如有个神灯精灵可以告诉你, 所有客人喝的汽水罐数的平均数或者是中位数二者之一。你会要求平均数还是中位数?为什么?再让答案可以更具具体些, 假设一共会有 30 位客人, 而精灵会告诉你 $\bar{x}=5$ 罐或者 $M=3$ 罐其中之一。这样你应该买多少罐汽水?

12.19 各州的 SAT 分数。我们想要比较美国各州 SAT 的数学平均分数以及语言平均分数的分布。我们把这些资料输入电脑, 用 SATM 代表数学分数, SATV 代表语言分数。以下是统计软件 Minitab 输出的结果。(用其他软件也可以得出近似的结果。有些软件计算四分位数的公式和我们用的有一点不一样。所以用软件得到的结果, 不一定和我们手算出来的完全一致。)

变量	<i>N</i>	平均数	中位数	标准差
SATV	51	532.22	525.00	33.86
SATM	51	533.47	528.00	35.13
变量	最小数	最大数	Q_1	Q_3
SATV	478.00	593.00	501.00	564.00
SATM	473.00	601.00	503.00	558.00

用这些输出的结果, 画出各州 SAT 数学分数箱形图和语言分数箱形图。用言语来叙述这两个分布有何不同。

12.20 SUV 费油吗?表 11.2 提供了 2000 年 32 型中型车的公路里程数(每加仑英里数)。你在习题 12.6 已经找出这些资料的五数综合。

以下是 26 种 2000 年的运动型多功能车(SUV)的公路里程数:

车型	每加仑英里数	车型	每加仑英里数
BMW X5	17	起亚运动型	22
雪佛兰布雷瑟	20	陆虎	17
雪佛兰塔荷	18	凌志 LX470	16
道奇杜兰哥	18	林肯领航员	17
福特远征者	18	马自达 MPV	19
福特探险者	20	奔驰 ML320	20
本田护照	20	三菱猎人	20
无限 QX4	18	日产开拓者	19



车型	每加仑英里数	车型	每加仑英里数
五十铃朋友	19	日产特锐	19
五十铃巡警	19	富士森林人	27
吉普切诺基	20	铃木大威达	20
吉普大切诺基	18	丰田RAV4	26
吉普牧马人	19	丰田赛跑者	21

- (a) 用图和数值来描述 SUV 的公路耗油情形。这个分布有些什么特点？
- (b) 画箱形图来比较中型车和 SUV 的公路耗油量。这两个分布最主要的差别有哪些？

12.21 热狗含多少卡路里？有些人会耽心自己究竟吃进多少卡路里。美国《消费者报告》杂志在一篇有关热狗的报导中，度量了 20 种品牌的牛肉热狗、17 种品牌的其他肉类热狗以及 17 种品牌家禽肉热狗的卡路里含量。

以下是牛肉热狗的电脑输出的资料：

平均数 = 156.8 标准差 = 22.64
最小数 = 111 最大数 = 190 N = 20
中位数 = 152.5 四分位数 = 140 178.5

再者是其他肉类热狗的资料：

平均数 = 158.7 标准差 = 25.24
最小数 = 107 最大数 = 195 N = 17
中位数 = 153 四分位数 = 139 179

而家禽肉热狗的资料为：

平均数 = 122.5 标准差 = 25.48
最小数 = 87 最大数 = 170 N = 17
中位数 = 129 四分位数 = 102 143

(有些软件用来找四分位数的公式和我们的有点不同。所以软件得出来的答案不见得和你手算的完全一样。)利用这些信息来选出三种热狗卡路里计数的箱形图。简略比较一下这三个分布。吃家禽肉做的热狗，比起牛肉热狗或其他肉类热狗，是不是通常会吃进较少的卡路里？



12.22 求标准差。血液中各种物质的含量多寡，会影响到我们的健康。以下是某位病人血液中磷酸盐的含量，单位是每十分之一升的血所含毫克数，这六个数字分别是病人每一次到诊所看病时所度量的：

5.2 5.2 4.6 47.9 5.7 6.4

只有 6 个观测值，画图也看不出什么名堂，所以我们就直接算平均数和标准差。

- (a) 利用平均数的定义来计算出平均数。也就是说，算出 6 个观测值的和，再除以 6。
- (b) 利用标准差的定义来算出标准差的值。也就是说，先找出每个观测值距平均数的距离，把这些距离平方，然后算出标准差。可以参考例 4。
- (c) 现在把你的数据输入计算机，再利用平均数键和标准差键来求得 \bar{x} 和 s 。这个结果和你手算的一样吗？

12.23 s 在度量什么。用计算机找出下列两组数字的平均数和标准差。

(a) 4 0 1 4 3 6。

(b) 5 3 1 3 4 2。

画一个并列的茎叶图(参考习题 11.16)来比较这两个分布。哪一个分散得比较开？

12.24 s 在度量什么。把上一习题(a)中的数字每一个加上 2。现在这组数字变成了 6 2 3 6 5 8。

- (a) 用计算机算出平均数和标准差，并且拿结果和上一题(a)的结果比一比。每一个数都加上 2，平均数会怎样变？标准差又会怎样变？
- (b) 如果我们把上一题(a)部分的数据，每一个都加上 10，则 \bar{x} 会变成怎样， s 又会变成怎样？不要做计算，直接回答。(这个习题是要说明一个事实，就是标准差度量的是，数据对应于其平均数的散布情形，若把整组数据同时移位，对标准差不会有影响。)

12.25 轿车和 SUV。用平均数和标准差来比较中型车(表 11.2)和 SUV(习题 12.20)的汽油里程。这些数字有没有体现出你在习题 12.20 所做较详细比较的要点？



12.26 来比赛。我们来做标准差竞赛。你必须从0到9之间的整数中间选出4个数，数字可以重复。

- (a) 选出4个数，使其标准差为最小。
- (b) 选出4个数，使其标准差为最大。
- (c) (a)和(b)还有没有其他正确答案。请说明。

12.27 光看 \bar{x} 和 s 还不够，平均数 \bar{x} 及标准差 s 分别度量中心和离度，但是并不能完整描述一个分布。分布形状不同的两组数据，有可能有相同的平均数和标准差。用你的计算机替次页上方的两小组数据找出平均数和标准差，来见证这个事实。然后替两组数据各画一个茎叶图，并评论两个分布的形状。

A 组数据	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74
B 组数据	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

12.28 调薪 某所学校给教师的年薪在30 000美元到60 000美元之间。教师会和学校的董事会正在协议次年的加薪幅度，假设最后给每位教师加薪1 000美元。

- (a) 平均薪水会增加多少？中位薪水会增加多少？
- (b) 如果离度用第一四分位数和第三四分位数之间的距离来度量的话，每人加薪1 000元会不会增加离度？
- (c) 如果用标准差来度量离度的话，每人加薪1 000美元会不会增加离度？(需要的话可参考习题12.24。)

12.29 调薪。假设上一题里的教师，每人下一年度的薪水都增加5%。所以薪水增加的幅度由1 500美元到3 000美元不等，视每人目前的薪水而定。如果用两个四分位数之间的距离来度量离度的话，每人加薪5%会不会增加离度？你觉得标准差会增加吗？

12.30 隐恶扬善。美国大专院校会宣布他们入学新生的“平均”SAT分数，而通常每所学校都希望这个“平均”愈高愈好。《纽约时报》一篇报道指出：“用奖学金来“大量收买”顶尖学生的私立学校喜欢用平均数，而谁都可以申请入学的公立学校喜欢用中位数。”运用你对于平均数和中位数的知识，来说明为什么私校和公立学校会各自有如此偏好。



12.31 要画什么图?我们已经学了三种可以用来展示数值变量分布的图:直方图、茎叶图及箱形图。举例说明(用话说明即可,不要提出数据)在什么情况下用哪种图最适合。

第 13 章

正态分布

科技行动

柱状图和直方图自然是很古老的东西了。用柱状图来呈现数据的历史，可以一直追溯到英国经济学家普莱费尔 (William Playfair, 1759—1823) 这位数据制图学 (data graphics) 的先驱。画直方图必须先选择如何分组，而不同的分组会呈现不同的图形。现代的软件如此发达，必定可以提供更好的方法来画分布吧？

利用软件，可以把直方图里的各个长方形以一条平滑的曲线取代，这条曲线代表分布的整体形状。看一看图 13.1，该图代表的数字是 1992—1996 年之间，美国 115 所大学中少数族裔学生获得工程博士的人数。我们在第 11 章见过这些数据，而图 13.1 里面的直方图

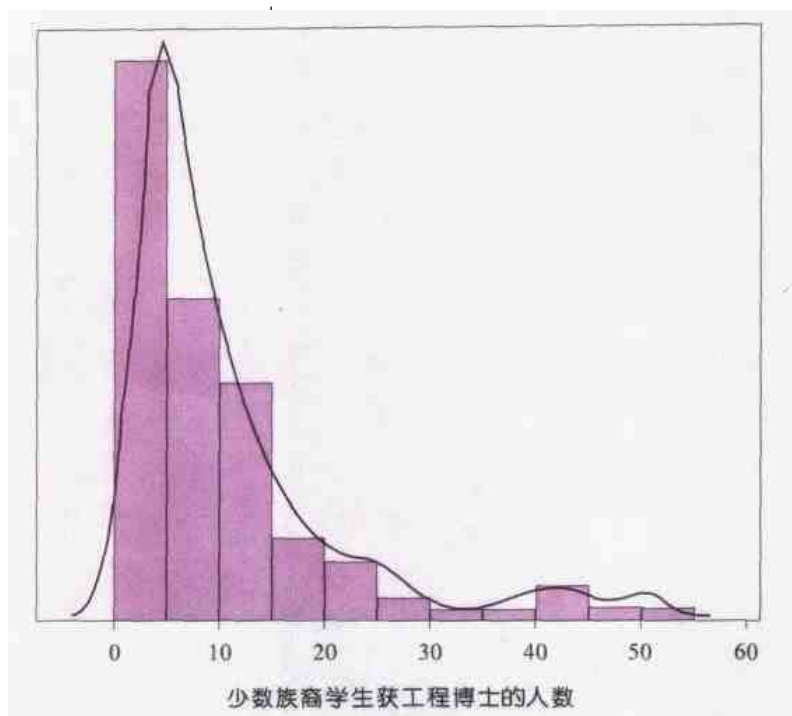


图 13.1 直方图和电脑绘出的曲线。图和曲线都是在描述 115 所大学中，少数族裔学生获工程博士人数的分布。这个分布是右偏的

和图 11.9 里的一样。图 13.1 里面的曲线，就是直方图在新科技之下的替代品。不过软件并不是根据直方图来画这条曲线的，你给它原始数据，它就会很聪明地画出这条曲线来描述分布。

在图 13.1 里，软件描绘了整体的形状，而且比直方图更有效的表达出右边尾巴处的波动状况。然而最高峰的表达却稍微有点困难，比如说，软件把曲线左端延伸过 0 的左边，以便使很尖锐的高峰稍微平顺些。在图 13.2 里，我们把同样的软件用在较大的一组数据上面，这组数据的分布形状比较有规则。这些数据是 1 000 个大小为 1 523 的简单随机样本 (SRS) 的样本比例 \hat{p} 的值，样本来自总体比例 $p=0.6$ 的总体。我们也在第 11 章中见过这些数据，而图 13.2 的直方图也是从图 11.4 复制过来的。软件画出的曲线展现的是一个特别对称且单峰的钟形。

虽然对于图 13.1 的不规则分布，我们没办法画出更好的曲线。然而对于图 13.2 这种很对称的抽样结果，却还有另一个方法可以得到平滑曲线。我们根据数学可以得知，这种分布可以用叫正态曲线 (normal curve) 的特殊平滑曲线来描述。图 13.3 画出来的曲线，就是针对这组数据所得到的正态曲线。这条曲线看起来很像图 13.2 里面



的那一条，然而仔细一点看的话会发现，这条曲线更为平顺些。正态曲线用起来很方便，也不需要用到聪明的软件。我们会看到，正态曲线有一些特别性质，让我们用起它来和考虑它的时候更为方便。不过只有某些类型的数据适合使用正态曲线，所以高科技软件还是要留着，以便在不适合使用正态曲线时拿出来用。

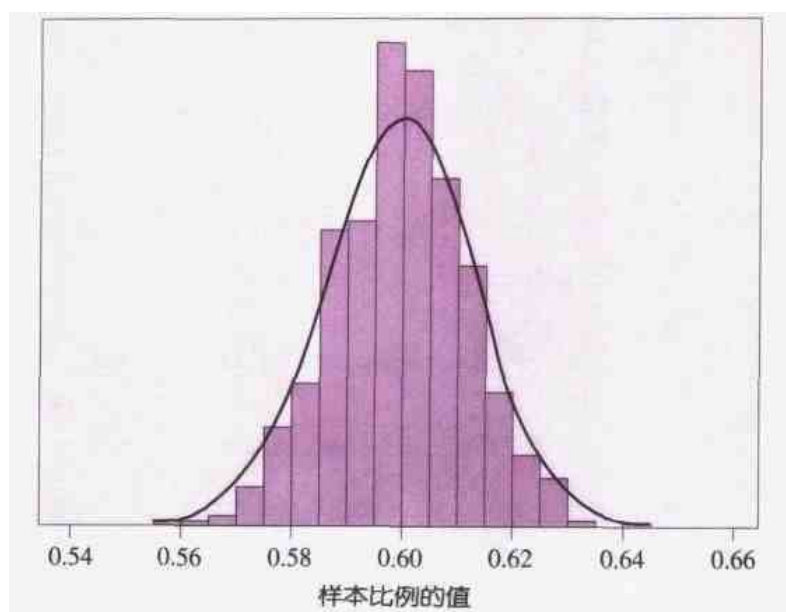


图 13.2 直方图和电脑绘出的曲线。二者全都是在描述从同一总体取出的 1 000 个简单随机样本的样本比例。这个分布相当对称

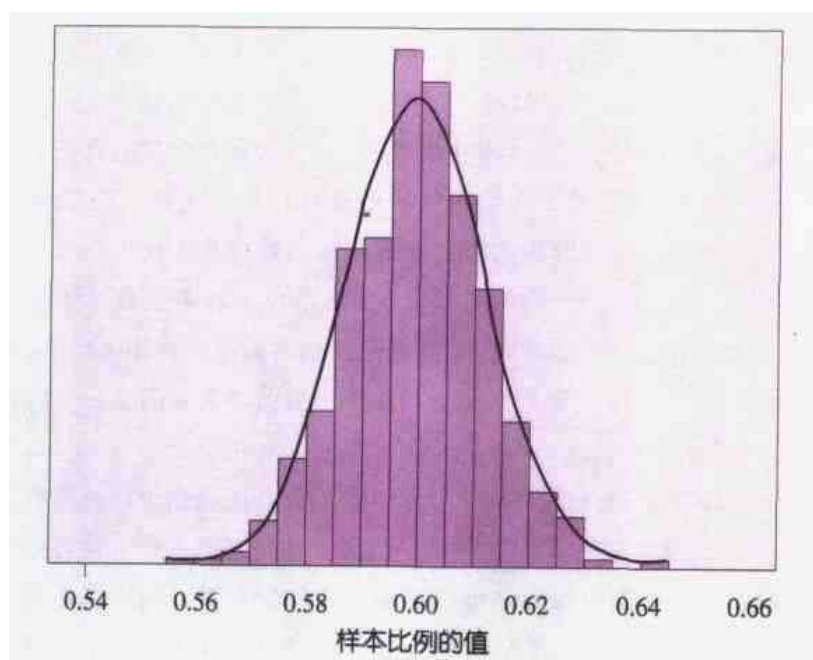


图 13.3 完全对称的正态曲线，用来描述样本比例的分布



我们现在有一整箱的工具可以来描述分布了，其中有图形，也有数值。还不止这样呢，对于探索单一数量变量的分布，我们也有一套明明白白的策略。

1. 一定要把数据画出来：通常是画直方图或茎叶图。
2. 寻找整体形态(形状、中心及离度)以及比如像异常值这样的显著偏差。
3. 选择用五数综合或者平均数和标准差来简略描述中心及离度。
4. 还可以给以上策略再加上一招：有时观测值数量多时，整体形态会显示出某种规律，即可以用平滑曲线来描述。

密度曲线

图 13.1 和 13.2 里显示了曲线如何代替直方图，来描绘数据分布的整体形状。你可以想像画一条曲线，穿过直方图里各长方形的顶

例 1 如何使用密度曲线

图 13.4 是从图 13.3 复制过来的，图的内容包括描述 1 000 个样本比例的直方图和正态曲线。其中比 0.61 大的观测值占怎样的比例？从 1 000 个实际观测值里去数的话，可以数出来共有 195 个观测值超过 0.61，所以它所占的比例是 $195/1\,000$ ，即 0.195。因为 0.61 在直方图里正好是在相邻两组的分界点上，所以图 13.4(a) 的斜线区域面积，就占全部长方形总面积的 0.195。

现在把焦点放在穿过直方图的密度曲线上。这个曲线底下的总面积是 1。而图 13.4(b) 中的斜线区域代表大于 0.61 的观测值所占比例，这个面积是 0.208。而 0.208 距 0.195 相当近，所以可看出密度曲线是很不错的近似方法。

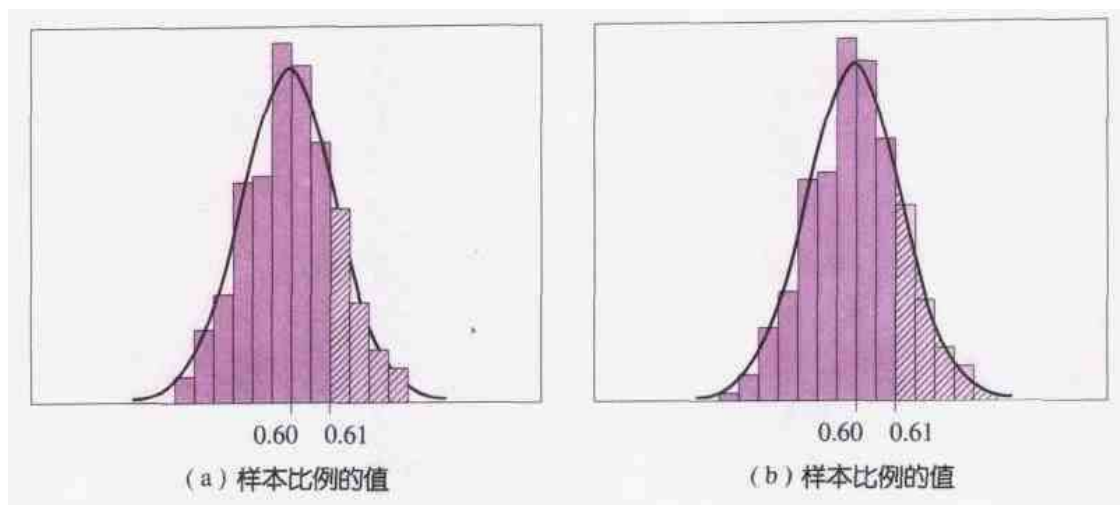


图 13.4 正态曲线及直方图。(a)直方图中的斜线区域的面积代表大于 0.61 的观测值。这在 1 000 个观测值中占了 195 个。(b)正态曲线之下的斜线面积代表大于 0.61 的观测值比例。此面积为 0.208

部，把长方形很不规则的高高低低给缓和掉。直方图和这些曲线之间有一个重要的差别。大部分的直方图是用长方形的高度来显示落在每组的观测值个数，也可以说是用长方形的面积来显示这些计数。我们画曲线的时候，它的结构都是利用曲线底下的面积来表示落在该区的观测值的比例(proportion)。为了做到这点，我们会选择适当尺度(scale)，使得曲线底下的总面积恰恰是 1。这样就会得到一个密度曲线(density curve)了。

因为密度曲线是把分布加以理想化之后所产生的图形，所以例 1 当中密度曲线底下的面积，和真正的比例并不相等。举例来说，曲线是完全对称的，但实际数据只是大致上对称。因为密度曲线是把分布的整体形状弄平滑之后的理想情况，所以对于描绘大量观测值的时候最为有用。

密度曲线的中心和离度

密度曲线可以帮我们进一步了解中心和离度的量度。中位数和四分位数很容易找。密度曲线底下的面积，代表占全体观测值的比例。中位数是左右各有一半观测值的那个点。所以一个密度曲线的中位数

就是等面积点(equal-areas point),也就是曲线底下的一半面积在它左边,另一半在它右边的那个点。四分位数把曲线底下的面积分成四等份。曲线底下四分之一的面积在第一四分位数的左边,四分之三的面积在第三四分位数的左边。用目测法把曲线底下的面积分成四等份,就可以大致找到任何密度曲线的中位数和四分位数。

因为密度曲线是理想化的形态,所以对称的密度曲线是百分之百对称的,因此对称密度曲线的中位数就在正中间。图 13.5(a)里就画出了对称曲线中位数的位置。而像图 13.5(b)里的偏斜曲线,我们也可以用目测方式大致找出等面积点。

平均数又如何呢?一组观测值的平均数就是它们的算术平均。如果我们把观测值想像成是叠在翘翘板上的砝码,平均数就是翘翘板的平衡点。而这点对于密度曲线来说也是正确的。假如密度曲线的图形是用实心材料做成的话,那么它的平衡点就是它的平均数的所在。图 13.6 说明了关于平均数的这项事实。对称的曲线因为两侧完全一样,所以平衡点就在中心位置。而对称的密度曲线,平均数和中位数刚好相等,如图 13.5(a)所显示的。我们知道偏斜分布的平均数会被拉往长尾方向。图 13.5(b)显示出,偏斜密度曲线的平均数,如何比中位数更被拉往长尾方向。

正态分布

图 13.3 和 13.4 里面的密度曲线同属一种特别重要的曲线:正态

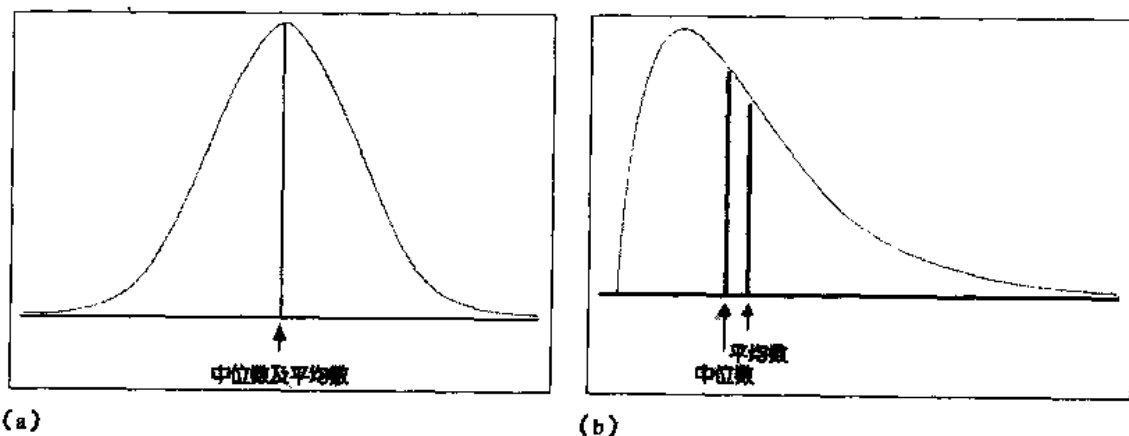


图 13.5 两个密度曲线的中位数及平均数,一为对称的正态曲线,一为右偏曲线



图 13.6 密度曲线的平均数就是它的平衡点

• 密度曲线的中位数和平均数

密度曲线的中位数是等面积点，也就是把曲线底面积分成两半的点。

密度曲线的平均数是平衡点。如果曲线是用实心材料做成的话，就会在那一点平衡。

对于一个对称的密度曲线来说，中位数和平均数是一样的。二者都在曲线的中心位置。而偏斜曲线的平均数会离开中位数，被拉向长尾方向。

曲线。图 13.7 再呈现了两个正态密度曲线(normal density curve)，正态曲线都是对称、单峰及钟形(bell-shaped)，尾部下降得很快，所以我们应该不会看到异常值。因为正态分布是对称的，所以平均数和中位数都落在曲线的中间位置，而这也是尖峰点所在。

正态曲线还有一个特别性质：我们可以用目测方式在曲线上找到它的标准差。对大部分其他的密度曲线，没有办法这样做。做法是这样的：想像你从山顶开始滑雪，山的形状和正态曲线一样。起先当你从山顶开始下滑时，往下的角度非常陡。

幸好，在你还没有直直坠下之前，斜坡便开始缓和起来，你愈往下滑出去，坡度愈平。



这个发生“曲率”(curvature)改变的地方,是在平均数两侧,各距平均数一个标准差的位置。在图 13.7 的两个曲线上都标示出了标准差。你如果用一支铅笔沿着正态曲线描,应该可以感受到曲率改变的地方,进而找出标准差。

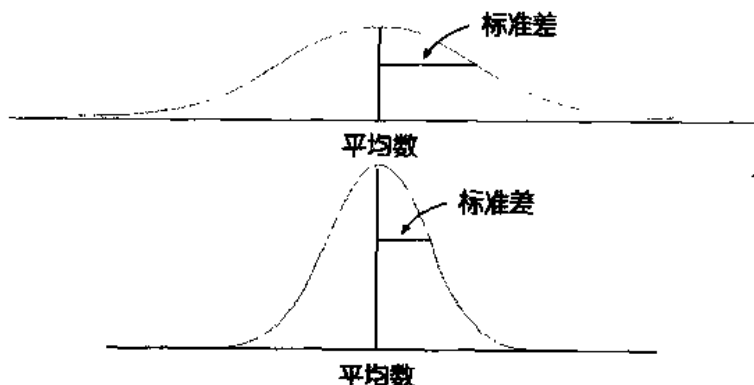


图 13.7 两个正态曲线。标准差决定了正态曲线的高度

正态曲线有个特别的性质:只要知道平均数及标准差,整个曲线就完全确定了。平均数把曲线的中心定下来,而标准差决定曲线的形状。变动正态分布的平均数并不会改变曲线的形状,只会改变曲线在 x 轴上的位置。但是,变动标准差却会改变正态曲线的形状,如图 13.7 所示。标准差较小的分布,散布的范围比较小,尖峰也比较陡。以下是正态曲线一些基本性质的总结。

• 正态密度曲线

正态曲线是对称的钟形曲线,具备以下性质:

- 只要给了平均数和标准差,就可以完全描述特定的正态曲线。
- 平均数决定分布的中心,这个位置就在曲线的对称中心(center of symmetry)。
- 标准差决定曲线的形状,标准差是指从平均数到平均数左侧或右侧的曲率转变点的距离。

为什么正态分布在统计里而很重要呢?首先,对于某些真实数据的分布,用正态曲线可以做很好的描述。最早将正态曲线用在数据上的是大数学家高斯(Carl Friedrich Gauss, 1777—1855)。天文学家或测



量员仔细重复度量同一个数量时，会有小误差，高斯用这些曲线来描述这些小误差。你有时候会看到有人把正态分布叫做“高斯分布”(Gaussian distribution)，就是为了纪念高斯。19世纪的大部分时间中，正态曲线叫做“误差曲线”(error curve)，因为正态曲线最早是用来描述量度误差的分布。后来慢慢发现，有些生物学或心理学的变量也至少大致是正态分布时，“误差曲线”这个名词就不再使用了。1889年时，高尔顿(Francis Galton, 1822—1911)创先把这些曲线称做“正态曲线”。高尔顿是达尔文(Charles Darwin, 1809—1882)的表弟，他开拓了遗传学的统计研究。

当我们从同一总体抽取许多样本时，诸如样本比例及样本平均数这类统计量的分布，也可以用正态曲线来描述。在图 13.3 和 13.4 中，我们就是这样在用正态分布的。抽样调查结果的误差界限，也常常用正态曲线来算。然而，即使有许多类的数据符合正态分布，仍然有许多是不符合的。比如说，大部分的收入分布是右偏的，而不是正态分布。非正态的资料就和平常的人一样，不仅常见，而且有时比正态的资料还有趣。

68 - 95 - 99.7 规则

正态曲线有许多，每一个正态曲线都可以用各自的平均数和标准差来描述。所有正态曲线都有许多共同性质，特别要提的是，对正态分布来说，标准差是理所当然的量度单位。这件事反映在下列规则当中。

◆ 68 - 95 - 99.7 规则

在任何正态分布当中，大约有：

- 68% 的观测值，落在距平均数一个标准差的范围内。
- 95% 的观测值，落在距平均数两个标准差的范围内。
- 99.7% 的观测值，落在距平均数三个标准差的范围内。

图 13.8 说明了 68 - 95 - 99.7 规则。记住这 3 个数字之后，你可以常常考虑到正态分布，却不用一直做啰唆的计算。不过还得记住，



没有哪组资料是百分之百用正态分布描述的。对于 SAT 分数, 或者蟋蟀的身长, 68-95-99.7 规则都只是大体正确。

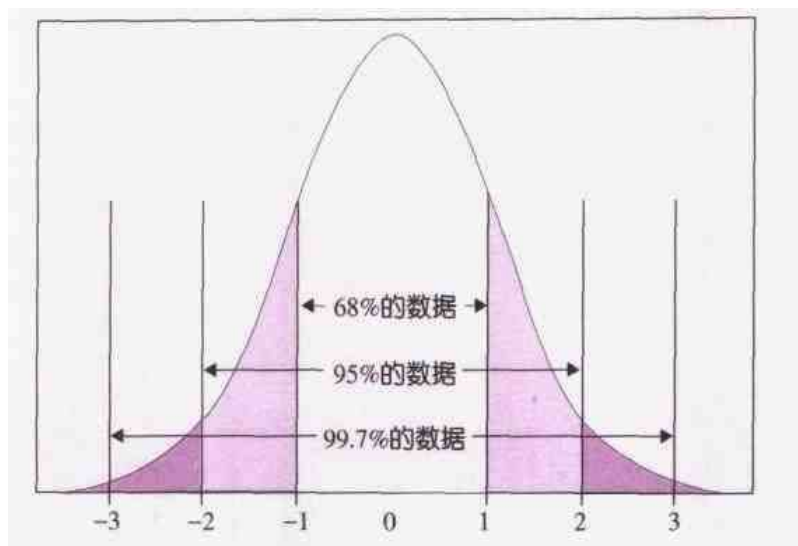


图 13.8 正态分布的 68-95-99.7 规则

例 2 年轻女性的身高

18—24 岁女性的身高, 约略是平均数 65 英寸、标准差 2.5 英寸的正态分布。要运用 68-95-99.7 规则, 首先得画一个正态曲线的图。图 13.9 说明了这个规则用在女性的身高上会是什么情况。

任何正态分布都有一半的观测值在平均数之上, 所以年轻女性中有一半高于 65 英寸。

任何正态分布的中间 68% 观测值, 会在距平均数一个标准差的范围内。而这 68% 中的一半, 即 34%, 会在平均数之上。所以有 34% 的年轻女性, 身高在 65—67.5 英寸之间。把身高不到 65 英寸的 50% 女性也加上, 可以得知总共有 84% 的年轻女性身高不到 67.5 英寸。所以推知超过 67.5 英寸的人占 16%。

任何正态分布的中间 95% 的值, 在距平均数两个标准差范围内。这里的两个标准差是 5 英寸, 所以年轻女性身高的中间 95% 是在 60 英寸(从 $65 - 5$ 得来的)和 70 英寸($65 + 5$)之间。

另外 5% 年轻女性的身高, 就超出 60—70 英寸的范围之外。因为正态分布是对称的, 其中有一半的女性是在矮的那一头。年轻女性中最矮的 2.5%, 身高不到 60



英寸(5英尺)。

任何正态分布中几乎所有(99.7%)的值,在距平均数三个标准差的范围内,所以几乎所有年轻女性的身高,都在57.5—72.5英寸之间。

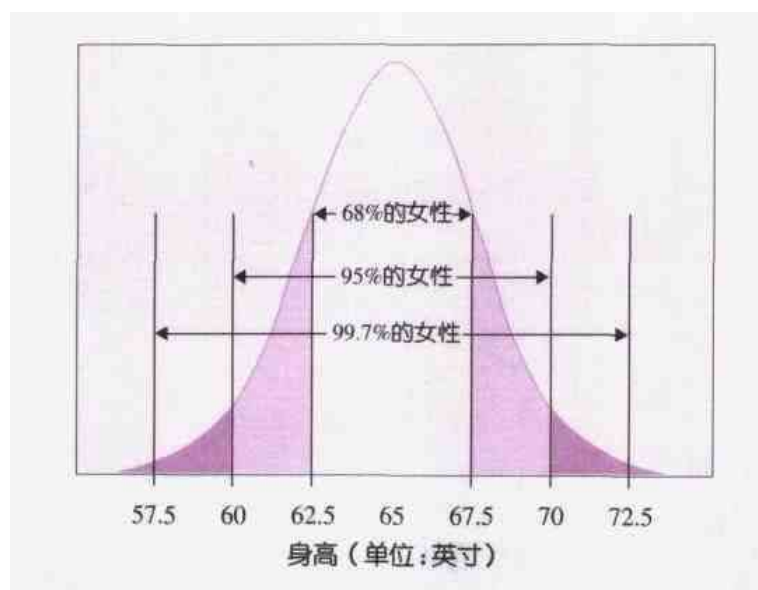


图 13.9 女性身高的 68-95-99.7 规则。这个正态分布的平均数为 65 英寸,标准差为 2.5 英寸

标准计分

珍妮在 SAT 大学入学测验的语言部分考了 600 分。这个成绩算不算好,这得要看在所有分数的分布中,600 分居于什么位置而定。SAT 测验经过规划,分数大致遵循平均数为 500、标准差为 100 的正态分布。珍妮的 600 分比平均数高上一个标准差。现在用 68-95-99.7 规则,就可以知道她到底考得怎样(图 13.10)。有一半考生的分数低于 500 分,另有 34% 在 500—600 分之间。所以珍妮比参加 SAT 测验的考生中的 84% 考得好。她的成绩报告上面不仅会列出她考了 600 分,还会加上说明,这个分数是第 84 百分位数(percentile)。这是



“比84%的考生考得好”的统计说法。

因为标准差是正态分布最自然的量度单位，所以我们可以换个方式，把珍妮的分数说成是“高于平均数一个标准差”。像这样以分布的平均数为“标兵”，而把观测值以距平均数几个标准差的方式表达出来，叫做标准计分(standard score)。

• 标准计分

任何观测值的标准计分(standard score)为：

$$\text{标准计分} = \frac{\text{观测值} - \text{平均数}}{\text{标准差}}$$

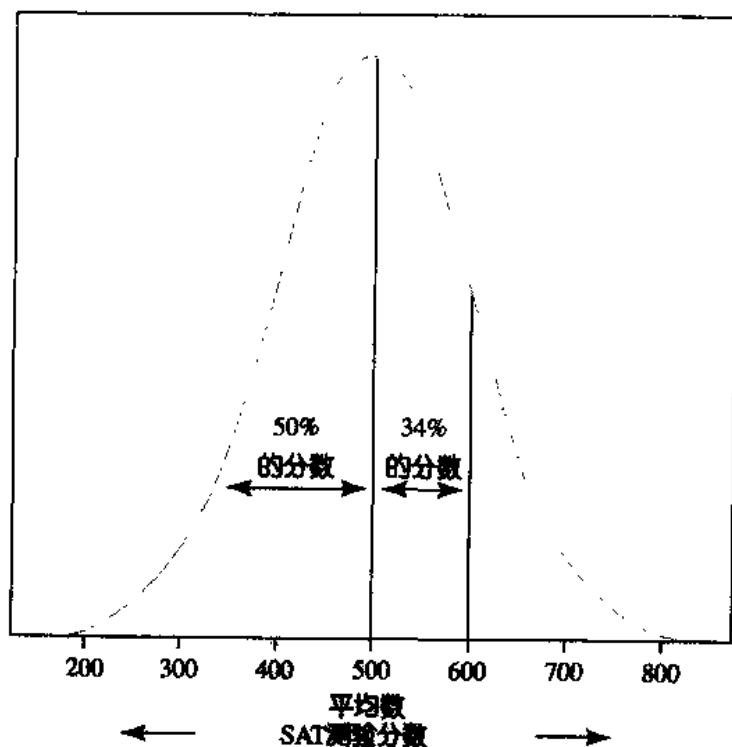


图 13.10 用 68-95-99.7 规则可以得出，任一正态分布都有 84% 的观测值在平均数以上一个标准差位置的左边。这个图里是把她的这个结果用在 SAT 分数上



是钟形曲线吗?

人的智慧高低的分布,是不是遵循正态分布的“钟形曲线”? IQ 测验的分数的大致符合正态分布。但那是因为测验分数是根据作答者的答案计算出来的,而计算方式原本就是以正态分布为目标设计的。要说智慧分布遵循钟形曲线的前提是:大家都同意 IQ 测验分数可以直接度量人的智慧。然而许多心理学家都不认为世界上有某种人类特征,可以称之为“智慧”,并且可以用一个测验分数度量出来。

标准计分为 1 的意思是说:所对应的观测值,在平均数之上一个标准差的位置。观测值的标准计分为 -2,就表示该观测值在平均数之下距离两个标准差的地方。标准计分可以用来比较不同分布中的值。当然,如果你不愿意用标准差来描述分布的高度,就不应该用标准计分。也就是说,分布必须至少是大致对称的,标准计分才适用。

例 3 ACT 分数对应 SAT 分数

珍妮在 SAT 的语言部分得了 600 分。她的朋友杰拉尔德参加了美国大学测验 (ACT, American College Testing), 在语言部分拿了 21 分。ACT 的分数是正态分布的, 平均数为 18, 标准差为 6。假设这两种测验的评量标准差不多, 谁的分数比较高?

珍妮的标准计分是:

$$\frac{600 - 500}{100} = \frac{100}{100} = 1.0$$

我们来跟杰拉尔德比一比, 他的标准计分是:

$$\frac{21 - 18}{6} = \frac{3}{6} = 0.5$$

因为珍妮的分数比平均高了 1 个标准差, 而杰拉尔德的分数只比平均高 0.5 个标准差, 所以珍妮考得比较好。

正态分布的百分位数*

对正态分布来说, 标准计分可以直接转换成百分位数, 而其他分布就没法子这样。

* 略过本节不会影响对本书其他部分的了解。



• 百分位数

一个分布的**第 c 百分位数**(the c th percentile)是一个值,指的是:小于第 c 百分位数的观测值,在全部观测值所占百分比为 c ,而其余的观测值则都比第 c 百分位数大。

任何分布的中位数就是分布的第 50 百分位数,而四分位数是第 25 及第 75 百分位数。在任何正态分布中,在平均数之上一个标准差的那点(标准计分为 1)是第 84 百分位数。从图 13.10 可看出为什么。正态分布的每个标准计分,都可以转换成特定的百分位数,而不论原来的正态分布的平均数和标准差是多少,所得百分位数都是一样的。本书末的表 B 中,列出对应于不同标准计分的百分位数。比起用 68-95-99.7 规则,用这个表可以做更多的细节计算。

例 4 大学入学测验的百分位数

珍妮 600 分的 SAT 分数可以转换成标准计分 1。我们已经知道根据 68-95-99.7 规则,这就是第 84 百分位数。而表 B 更精确些,它说标准计分 1 是正态分布的 84.13 百分位数。杰拉尔德在 ACT 拿的 21 分等于标准计分 0.5。表 B 说这是 69.15 百分位数。杰拉尔德考得不错,但没有珍妮好。百分位数比原始分数或标准计分都更容易了解。这就是为什么像 SAT 这类测验的成绩单上,通常都同时列出分数及百分位数。

例 4 找出对应某百分位数的观测值

学生要在 SAT 考多高的分数,才能跻身最高的 20% 呢?这个分数必须至少等于第 80 百分位数。到表 B 的内部去找最接近 80 的百分位数。你会看到,标准计分 0.8 对应 78.81 百分位数,而标准计分 0.9 对应 81.59 百分位数。表里面离 80 最接



近的百分位数是 78.81, 所以我们可以下结论, 对任何正态分布来说, 标准计分 0.8 大约等于第 80 百分位数。

要把标准计分还原为 SAT 分数, 只要把计算标准计分的步骤“倒过来”即可, 方法如下:

$$\begin{aligned}\text{观测值} &= \text{平均数} + \text{标准计分} \times \text{标准差} \\ &= 500 + (0.8)(100) = 580\end{aligned}$$

考到 580 分以上(含 580), 就会在最高 20% 范围内了。(更确切一点说的话, 这些分数属于最高的 21.19%, 因为 580 事实上是第 78.81 百分位数, 但是我们只要在表 B 里找最接近的数字就够了。)

网络寻奇

你见过列出数字平方根的表吗? 以前这种表很常见, 但是现在已经被计算机上的 \sqrt{x} 键给取代了。本书末的表 B 列出正态曲线底下的面积, 这种列表方式虽还是很常见, 但是也开始让路给一些按键, 甚至小程序了; 小程序还让你看到面积如何变化。请访问《统计学的世界》原文版的网站, www.whfreeman.com/scc, 并点选正态曲线小程序(“Normal curve” applet)。你甚至还可以利用这个小程序来做那些需要计算正态曲线面积的习题呢。



本章重点摘要

茎叶图、直方图和箱形图全都可以用来描述数值变量的分布。密度曲线是另一种图，但也做同样用途。密度曲线底下的面积必定是 1，而曲线的形状可以描述一个分布的整体形态。曲线底下的面积，代表观测值会落在对应的区间内的比例。用目测法可以找到密度曲线的中位数(等面积点)及平均数(平衡点)的大致位置。

正态曲线是一种特别的密度曲线，适合用来描述某些种类数据的整体形态。正态曲线是对称的钟形。特定的正态曲线可以完全由它的平均数和标准差来决定。你可以从正态曲线上找出平均数(中间点)的位置以及标准差的大小(从平均数到曲率改变的点之距离)。所有正态分布都遵循 **68 - 95 - 99.7** 规则。标准计分是以标准差为单位，把观测值表示成距离平均数有几个单位，平均数的标准计分是 0。标准计分所对应的百分位数，在所有正态分布都是一样的。表 B 列出了正态分布的百分位数。



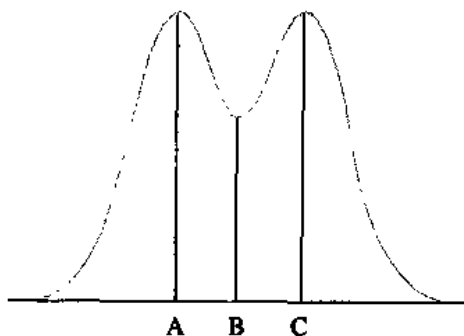
第13章 习题

13.1 密度曲线。

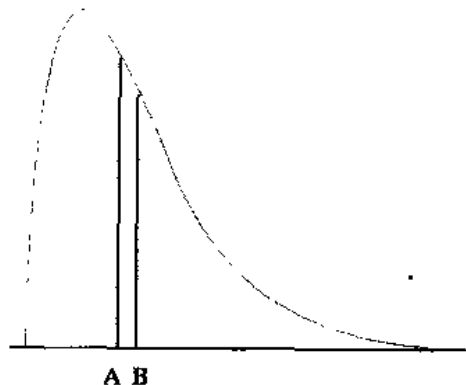
- (a) 画一条对称的密度曲线，但是形状和正态曲线不同。
- (b) 画一条非常左偏的密度曲线。

13.2 平均数和中位数。图 13.11 里有好几条形状不同的密度曲线。大致描述一下每个分布的整体形状。在图里面有标示出的一些点，而平均数和中位数就在这些点之中。对每个曲线找出哪一点是中位数，哪一点又是平均数。

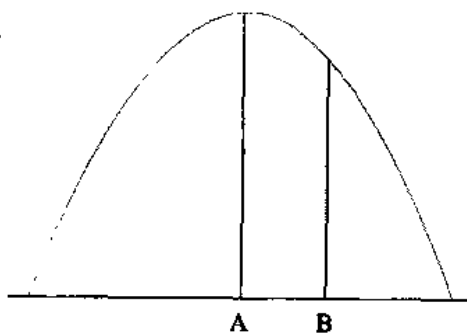
(a)



(b)



(c)



(d)

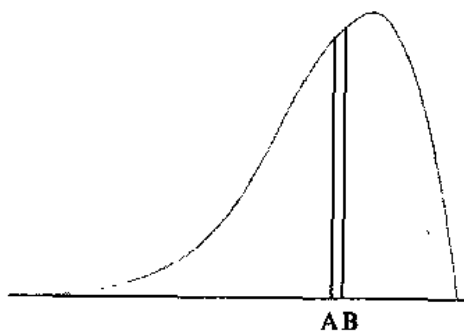


图 13.11 不同形状的 4 条密度曲线，对照习题 13.2。每条曲线的平均数和中位数，都在标示出的点中



13.3 随机数字 如果你指示电脑送出 0—1 之间的“随机数字”(random number), 你就会得到均匀分布(uniform distribution)的观测值。图 13.12 里画出了均匀分布的密度曲线。这条曲线在 0—1 之间的值都是 1, 而在这个范围之外都等于 0, 用这条密度曲线来回答以下问题。

- (a) 曲线底下的面积为什么等于 1?
- (b) 曲线是对称的。平均数以及中位数会等于多少?
- (c) 观测值中有多少百分比落在 0—0.4 之间?

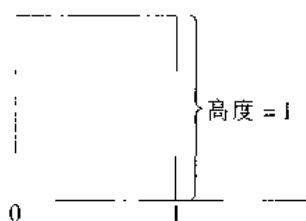


图 13.12 均匀分布的密度曲线, 对照习题 13.3。这个分布的观测值在 0—1 之间“随机”散布

IQ 测验分数。图 13.13 是 74 位七年级*学生 IQ 测验分散的茎叶图。这个分布很接近平均数为 111, 标准差为 11 的正态分布。这些学生包括美国中西部一所乡间学校的全部七年级生, 除了 4 个例外; 这 4 名学生不是生病, 就是因故没专心考试。得到的分数是超低的异常值, 因此被去除。我们就用平均数 111, 标准差 11 的正态分布, 当做所有中西部乡间学校七年级生 IQ 测验分数的分布。利用这个分布以及 68—95—99.7 规则, 来回答习题 13.4—13.6。

*译注: 相当于我们的初一。

13.4 所有中西部乡间学校七年级生当中 95% 的人, IQ 分数会介于哪两个值之间?

13.5 中西部乡间学校七年级生当中, 分数超过 100 的占多少百分比?

13.6 IQ 分数超过 144 的学生, 占多少百分比? 我们样本学校中的 74 个学生, 没人考这么高分。你对这个结果惊讶吗? 为什么?



8	6 9
9	0 1 3 3
9	6 7 7 8
10	0 0 2 2 3 3 3 3 4 4
10	5 5 5 6 6 6 7 7 7 7 8 9
11	0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 4 4 4 4
11	5 5 6 8 8 9 9 9
12	0 0 3 3 4 4
12	6 7 7 8 8 8
13	0 2
13	6

图 13.13 74 位七年级学生 IQ 测验分数的茎叶图, 对照习题 13.4—13.6

13.7 怀孕期的长短。人类从受孕到分娩的怀孕期, 长短各有不同, 但大致遵循平均数 266 天, 标准差 16 天的正态分布。用 68-95-99.7 规则回答下列问题。

- (a) 中间 95% 的怀孕期会落在哪两个数字之间?
 (b) 怀孕期最短的 2.5% 会有多短?

13.8 正态曲线。图 13.14 中的正态曲线, 平均数和标准差各是多少?

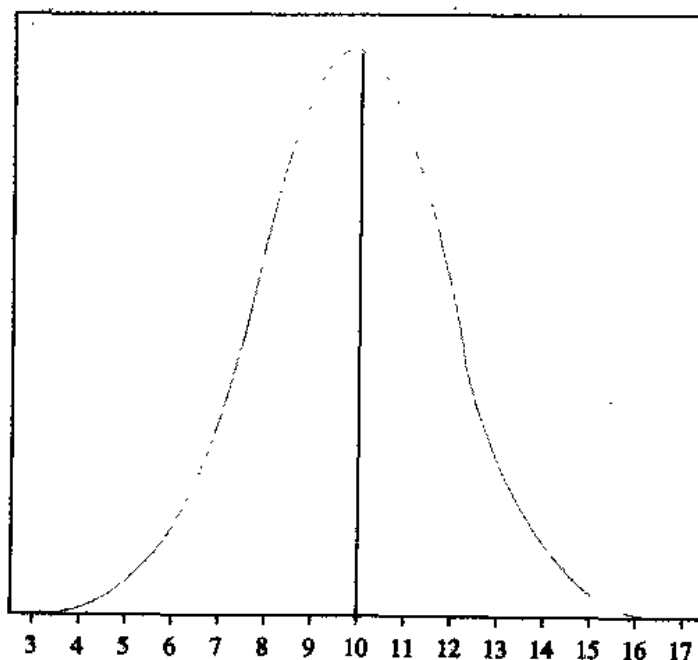


图 13.14 这个正态曲线的平均数和标准差各是多少? 见习题 13.8



13.9 马的怀孕期 大型动物通常怀孕的时间比较长。雌马从受孕到生出小马的怀孕期长度,大致遵循正态分布,平均数为336天,标准差3天。用68-95-99.7规则回答下列问题。

- (a) 几乎所有的(99.7%)雌马怀孕期的长度,都会在那个范围?
- (b) 有多少百分比的马,怀孕超过339天?

13.10 三位强打者 美国棒球史上具有划时代意义的三项纪录,一是科布(Ty Cobb)在1911年所创造的0.420平均打击率,一是泰德威廉斯(Ted Williams)在1941年的0.406,以及布雷特(George Brett)在1980年的0.390。我们不能直接比较这些平均打击率,因为大联盟平均打击率的分布逐年在改变,这些分布相当对称,而且大致接近正态曲线(除了科布,泰德威廉斯及布雷特这几个异常值以外)。平均打击率的平均数,多年以来因为规则改变以及投、打之间能力的平衡,而大致维持不变,且标准差逐渐下降。以下是实际数字:

年代	平均数	标准差
1910	0.266	0.037 1
1940	0.267	0.032 6
1970	0.261	0.031 7

分别计算一下科布、威廉斯及布雷特平均打击率的标准计分,来比较一下,每个人和同行比起来到底有多突出。

13.11 比较 IQ 分数 韦氏成人智力量表(WAIS, Wechsler Adult Intelligence Scale)是一种“IQ 测验”。20—34 岁年龄组的 WAIS 分数大致是正态分布,平均数为110,标准差25。60—64 岁年龄组的分数也大致是正态分布,平均数为90,标准差25。30 岁的莎拉在 WAIS 得到的分数是135。她60 岁的母亲也做了这个测验,得了120 分。

- (a) 把两个分数分别表示成标准计分,从而可看出两位女士在各自年龄组中居于什么位置。
- (b) 相对于自己的年龄群,莎拉和她母亲谁的分数比较高?对这个测验所度量的变量来说,她们二位谁的程度比较高?

13.12 男性的身高 年轻男性的身高分布,大致是平均数70英



寸、标准差 2.5 英寸的正态分布。画出这条正态曲线，平均数和标准差要在正确的位置。(提示：先画出曲线，找出曲率改变的地方，然后在横轴上标示出对应的位置。)

13.13 再谈男性身高。年轻男性的身高分布，大致是平均数 70 英寸，标准差 2.5 英寸的正态分布。用 68-95-99.7 规则来回答下列问题。

- (a) 有多少百分比的男性高于 75 英寸？
- (b) 最中间 95% 的男性，身高会介于哪两个高度之间？
- (c) 身高不到 67.5 英寸的男性占多少百分比？

13.14 男性和女性的身高。年轻女性的身高，大致是平均数 65 英寸，标准差 2.5 英寸的正态分布。同年龄层的男性身高，平均数(及中位数)是 70 英寸。请问有多少百分比的女性，比中等身高的男性要高？

13.15 年轻人的身高。18—24 岁美国男性的平均身高约为 70 英寸。同年龄层的美国女性平均身高约为 65 英寸。你觉得所有 18—24 岁美国人的身高分布，会不会接近正态分布？说明你的答案。

13.16 抽样。假设成年美国人中，因安全理由夜间不敢出门的比例是 $p = 0.4$ 。如果我们取很多个大小为 1 050 的 SRS，样本比例 \hat{p} 的值会随着样本而变，但遵循平均数 0.4、标准差 0.015 的正态分布。用这个事实，加上 68-95-99.7 规则，来回答下面问题。

- (a) 在众多样本当中， \hat{p} 的值有多少百分比会超过 0.4？超过 0.43 的又有多少？
- (b) 在大量的样本当中，中间 95% 的 \hat{p} 值会在什么范围？

13.17 我会当选吗？美国国会议员佛洛埃对选民做了一项抽样调查，以了解有多少百分比的选民支持他竞选连任。为了省钱，他只抽出 400 人的样本。假设实际上有 45% 的选民支持佛洛埃。大小为 400 的随机样本中支持佛洛埃的百分比，会随着样本而变，其分布遵循平均数 45%、标准差 2.5% 的正态分布。所有这样大小的样本中，有多少百分比会显示出超过一半选民支持佛洛埃的(与事实不符的)结果？



以下的额外习题，需要用到表 B 的正态分布百分位数。

13.18 NCAA 运动员规则。美国的全国大学生体育协会(NCAA, National Collegiate Athletic Association)要求第一类(Division One)运动员，在 SAT 测验的数学及语言部分，总共至少要得 820 分，才能在大学一年级参加竞赛。(对高中成绩很差的学生，所要求的分数还更高些。)1999 年的时候，总共数百万学生考出来的 SAT 分数，大致符合平均数 1017、标准差 209 的正态分布。其中分数低于 820 的学生，占多少百分比？

13.19 其他 NCAA 规则。若一个学生的 SAT 分数至少有 720 分的话，NCAA 认为他是“部分合格”，可以参与训练及接受运动奖学金，但不能参加竞赛。利用上一习题所给的资料来计算一下，SAT 分数低于 720 的占多少百分比。

13.20 SAT 800 分。在 SAT 考试的两个部分都可能超过 800 分，只是超过 800 在成绩单上也只记录成 800 分而已。(也就是说，学生拿到 800 分的成绩单，不代表他考了满分。)1999 年时，男性在 SAT 数学部分的分数遵循正态分布，平均数 531、标准差 115。有多少百分比的分数是超过 800(因此记录成 800)的？

13.21 女生的 SAT 分数。女生在 SAT 的平均表现，尤其是数学部分，比男生要差。对于这个性别差异背后的原因，大家各有不同意见。1999 年的女生数学 SAT 分数，遵循平均数 495、标准差 109 的正态分布。男生的平均数是 531。请问有多少百分比的女生，分数高于男生的平均？

13.22 我们是不是愈来愈聪明？当斯坦福-比奈(Stanford-Binet)“IQ 测验”在 1932 年开始使用的时候，曾经做过调整使得儿童的各年龄层的分数都大致符合平均数为 100，标准差为 15 的正态分布。之后该测验仍然不时做调整，以便使平均数保持在 100。如果叫当今的美国儿童去考 1932 年的斯坦福测验的话，他们的平均分数差不多会是 120。IQ 分数随着时间而愈来愈高的原因不得而知，可能有部分原因是与现代儿童营养较佳以及考试经验比较丰富有关。

(a) IQ 分数超过 130 通常叫做“超级优秀”。1932 年的超级优秀儿



童占多少百分比?

- (b) 如果现代儿童去考 1932 年的测验的话, 有多少百分比会考出“超级优秀”的分数(假设标准差没有变仍是 15)?

13.23 日本人的 IQ 分数 韦氏儿童智力量表(Wechsler Intelligence Scale for Children)在美国及欧洲都有人使用(它有好几种语言), 在每个地方的分数都大致是平均数 100、标准差 15 的正态分布。当在日本做调整之前, 平均数是 111。日本的这个平均数, 对应于美国-欧洲分数分布, 会是什么百分位数?

13.24 股市。股价指数(由许多不同股票组合而成)的每年获利率, 还算接近正态分布。自 1945 年以来, 标准普尔 500 指数的平均年获利率是 12%, 标准差是 16.5%。我们就把长期以来的年获利率看成是平均数 12%、标准差 16.5% 的正态分布。

- (a) 中间 95% 的年获利率分布在哪个范围内?
(b) 若有一年的指数获利小于 0, 我们就说那一年的股市是收黑的。收黑的那些年, 总共占多少百分比?
(c) 全年获利超过 25% 的占多少百分比?

13.25 找出四分位数。一个分布的四分位数, 是指第 25 和 75 百分位数。对正态分布来说, 四分位数距平均数约几个标准差?

13.26 年轻女性的身高。18—24 岁的女性身高, 大致是平均数 65 英寸、标准差 2.5 英寸的正态分布。最高 10% 的女性有多高?(利用在表 B 中最接近的百分比来算。)

13.27 高 IQ。20—34 岁年龄层的人, WAIS 分数大致遵循正态分布, 平均数为 110, 标准差为 25。要考多高的分才可以跻身前 25% 的高 IQ 群?

第 14 章

描述相关关系的方法：散布图和相关系数

替各州排名次？

媒体有公布排名顺序的癖好。最适合居住的城市、最好的大学、最健康的食物、服装最差的男士……，只要是最佳或最差的排名清单，几乎一定会出现在新闻报道中。因此每一年当各州 SAT 分数出来的时候，新闻报道会依各州高中毕业班学生的平均 SAT 分数，从最好的州（艾奥瓦）一直排到最差的州（南卡罗来纳）时，也就不令人惊讶了。

主办 SAT 考试的大学委员会很不喜欢媒体这么做。“只依照 SAT 分数来给各州做比较或排序是没有意义的，大学委员会非常不鼓励这种做法。”在列出各州平均 SAT 分数的表一开头便如此说道。



要知道为什么如此, 让我们来看一下数据。

图 14.1 显示出美国 50 州加上哥伦比亚特区, SAT 测验数学平均分数的分布。在 SAT 最低 400 最高 800 分的范围内, 艾奥瓦州以平均 601 分夺魁, 而南卡罗来纳则以 473 分垫底。这个分布的形状有点特别: 它有两个明显的峰。双峰分布提供的线索, 指向数据可能是由两组不同数据混合而成。图 14.2 揭开了这层神秘面纱。这个图是一个散布图 (scatterplot), 它把每个州的平均 SAT 分数当做代表该州的点的纵座标, 而该点的横座标是该州高中毕业生中参加了 SAT 测验者所占比例。图里可清楚看出全部的州分成两群: 其中一群参加测验的学生不超过三分之一, 而该群的平均分数都较高; 另一群中的州, 有超过一半的学生考了 SAT, 而平均分数较低。从整个图可以看出, 参加测验的学生愈多, 平均分数就愈低。

事实上有两个大学入学测验是最普及的: SAT 和 ACT。有些州偏好前者, 有些州偏好后者。在采用 ACT 的州, 只有申请某些特定学校的学生才去考 SAT, 而他们通常考得很好。难怪只有 5% 高中毕业生考 SAT 的艾奥瓦州 (ACT 的大本营), 会比有 76% 学生考 SAT 的纽约州考得好。艾奥瓦州的高分, 不能代表该州的教育品质优于纽约州。

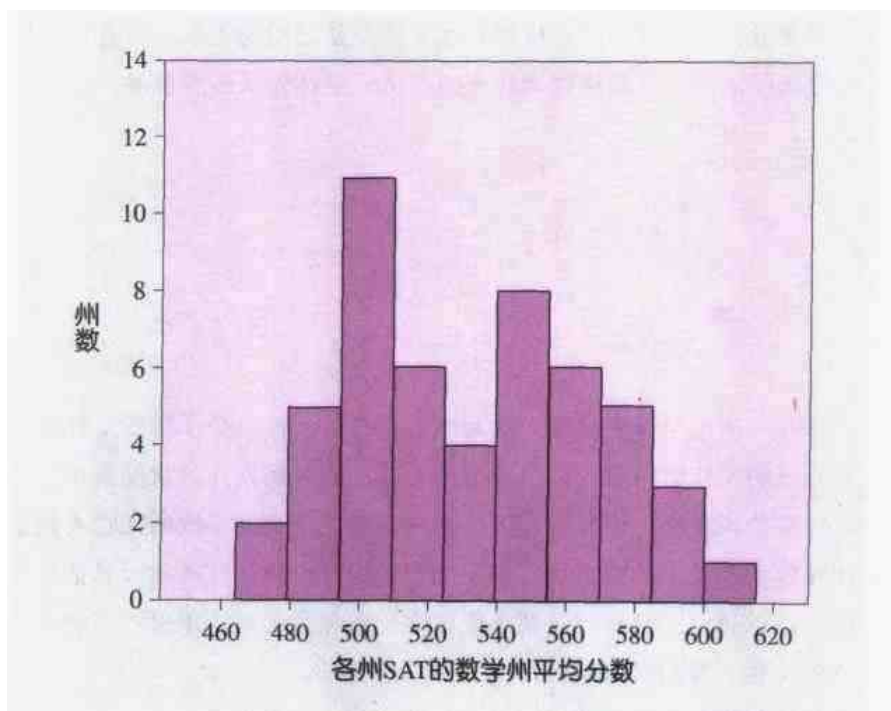


图 14.1 全美 50 州以及哥伦比亚特区的 SAT 测验数学平均分数的直方图

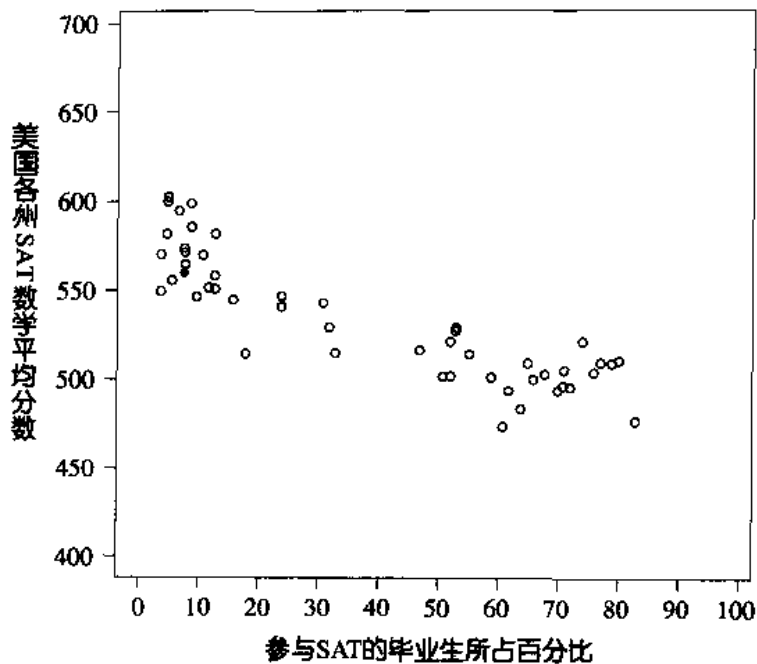
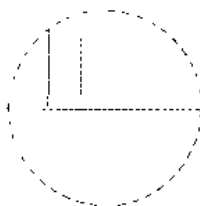


图 14.2 “美国各州 SAT 数学平均分数”对应“高中毕业生参加 SAT 测验百分比”的散布图

从我们对 SAT 分数的探讨，可以学到一般通用的道理：要了解一个变量，我们常常需要检视它和其他变量之间的关系。还有，我们也再次印证了：要知道数据代表的意义，应该先从画图着手。



有一项医学研究发现，比起中等身高的女性，个子矮的女性较常有心脏病发作的情形，而个子高的女性，心脏病发作的状况最少。某个保险公司宣称，以登记在案的每一万辆汽车交通事故的死亡人数来比较的话，较重的汽车死亡率要比较轻的汽车低。这两项及其他许多项的统计研究，都探讨过两个变量之间的相关关系。不过要了解这类的关联，我们常常还得检视一下其他变量。

比如说，如果想要做出“较矮的女性心脏病发作的风险较高”的结论，研究者首先必须能够消除掉其他诸如体重和运动习惯等变量的



影响。本章和接下来几章的主题,就是变量之间的相关关系。我们的重点之一,是两个变量之间的关系有可能受到一些隐藏起来的变量的重大影响。

大部分统计研究的数据都是对应不止一个变量的。幸运的是,对十多变量数据的分析,仍是以我们在研究单一变量时所用的工具为基础。分析时应遵循的原则和以前一样:

- 先用数据画图,并加入一些具代表性的综合数值。
- 寻找整体形态以及有异于整体形态的偏差。
- 当整体形态很有规则时,有时可以用很精简的方式来描述它。

散布图

最常用来展现二个数量变量之间的关系的是散布图。图 14.2 的散布图就显示出,各州平均 SAT 分数和参与 SAT 测验的高中毕业生百分比之间的相关性。我们认为“参加测验的百分比”有助于解释“平均分数”。也就是说,“参加测验的百分比”是解释变量(explanatory variable),而“平均分数”是反应变量(response variable)。

我们想知道当参加测验的百分比改变时,平均分数会有什么变化,所以把参加测验的百分比(解释变量)放在横轴上。结果我们看到当百分比上升时,平均分数就下降。图上面的每一个点代表一州。

举例来说,亚拉巴马州有 8% 的毕业生考了 SAT,而平均数学分数是 558。在 x 轴(横轴)上找出 8 的位置,并在 y 轴(纵轴)上找出 558 的位置。亚拉巴马就是在 8 的正上方以及 558 正右方的那个深色的点。

• 散布图

散布图(scatter plot)显示了在同一个个体上度量到的两个数量变量之间的关系。其中一个变量的值在横轴上标示,另一个变量的值在纵轴上标示。每一笔资料对应图中的一个点,点的位置由该个体两个变量的值决定。

• 如果有解释变量的话,一定要把解释变量标示在散布图的横轴(也就是 x 轴)上,提醒您一下,我们通常把解释变量叫做 x ,而把反



应变变量叫做 y 。而如果两个变量间没有“解释—反应”这样的差别，把哪个变量标示在横轴都无所谓。

例 1 健康与财富

图 14.3 的散布图是以来自世界银行(World Bank)的资料画成的。其中的个体是全世界每一个曾提供资料的国家，解释变量是国家富有程度的一种量度，即每人平均“国内生产总值”(GDP, gross domestic product)。GDP 是一个国家全部的生产和服务总值，通常都换算成美元表示。反应变量是每个人出生时的预期寿命。

我们会认为有钱的国家的人民应该活得长些。散布图的整体形态的确显示出这种状况，但是这里两个变量间的关联，出现有趣的形状。当 GDP 增加时，起先预期寿命急速增加，但是后来就拉平了。像美国这样富国的人民，并不会比稍贫穷但并非极穷国家的人民活得更久。其中有些国家，比如哥斯达黎加，甚至预期寿命还高于美国。

其中有三个非洲国家属于异常值。它们的预期寿命和邻国差不多，但是 GDP 比较高，分别是产油的加蓬，以及出产钻石的纳米比亚和博茨瓦纳。有可能因为出口矿产的收入主要进了少数人的口袋，因此把每人平均 GDP 拉高了，而对大部分老百姓的收入或者预期寿命却没什么帮助。也就是说，每人平均国内生产总值是一个平均数，而我们知道平均收入可能远高于中位收入。以博茨瓦纳来说，我们得到的 GDP 的资料还可能不是正确的。世界银行估计的每人平均 GDP 是 8 310 美元，图 14.3 用的是这个值，但是 CIA 的估计却只有 3 600 美元。

诠释散布图

想要解释散布图，用数据分析的一般方法即可。



• 检视散布图

在根据数据画的任何图里面, 要寻找**整体形态**以及明显偏离整体形态的**偏差**。

要描述散布图的整体形态, 可以描述点的**形式(form)**、**方向(direction)**及相关关系的**强度**。

有一种重要的偏差是**异常值**, 也就是落在相关关系的整体形态之外的个别值。

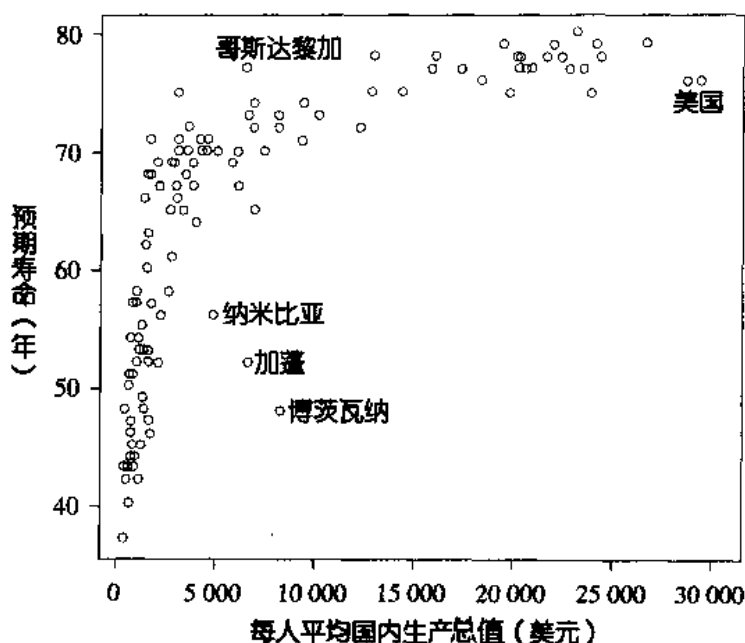


图 14.3 各国人民预期寿命对应于该国每人平均国内生产总值的散布图

图 14.2 和 14.3 都有明确的方向: 参加测验的学生愈多, 平均 SAT 分数就愈低; GDP 增加, 预期寿命大致上也增加。我们说图 14.2 显示出变量之间的负相关(negative association), 而图 14.3 显示出正相关(positive association)。

• 正相关与负相关

如果有两个变量, 当其中一个变量的值高于平均时, 另一变量的值也倾向高于平均, 而其中一个低于平均时, 另一个变量也倾向低于平均, 则称此二个变量是正相关的(positively associated)。此时散布图是从左到右往上斜的。

如果这两个变量的情况是: 一个变量的值高于平均时, 另一变量的值倾向低于平均, 前者低于平均时, 后者又倾向高于平均, 则称此二变量为负相关的(negatively associated)。此时散布图为从左到右往下斜。



我们的散布图都有显著的形式。图 14.2 显示出集结成两个群 (cluster) 的州, 而图 14.3 显示出曲线相关 (curved relationship)。散布图的相关强度, 是由图中的点与某个明确的形式有多接近而决定的。图 14.2 和 14.3 里的相关性不算强。参加 SAT 的百分比接近的某些州, 平均分数却颇有差距, 而 GDP 差不多的国家, 也可能有很不一样的预期寿命。以下的例子中, 有形式很简单但强度却较大的相关性。

例 2 将化石分类

始祖鸟 (archaeopteryx) 是一种已灭绝的动物, 它有像鸟类一样的羽毛, 但是也有像爬虫类的牙齿及长而多骨的尾巴。已知的化石标本只有 6 个, 因为这些标本的大小差很多, 有些科学家认为这些标本可能是不同的种类, 而不是同一种类的不同个体。在 5 个仍同时保有股骨 (一种腿骨) 以及肱骨 (上臂的骨头) 的标本中, 我们检查股骨及肱骨的长度, 以下就是这组资料, 单位是厘米:

股骨	38	56	59	64	74
肱骨	41	63	70	72	84

因为两个变量之间并没有“解释—反应”的差别, 我们的散布图中, 把哪个变量放在 x 轴都没关系。画出的图为图 14.4。

散布图显示出很强的正向直线相关。直线是重要的相关形式, 因为它既常见又简单好应用。相关性很强, 是因为点的分布很接近一直线。是正向的, 是因为当一种骨头的长度增加时, 另一种骨头的长度也增加。从这些资料看来, 这 5 件化石应该都属于同一种类, 而大小不一样, 是因为有的比较年轻。我们认为: 在这两种骨头的长度之间, 不同的种类应该有不同的相关性, 因此, 在这个散布图上, 一个不同的种类应该会对应一个异常点。

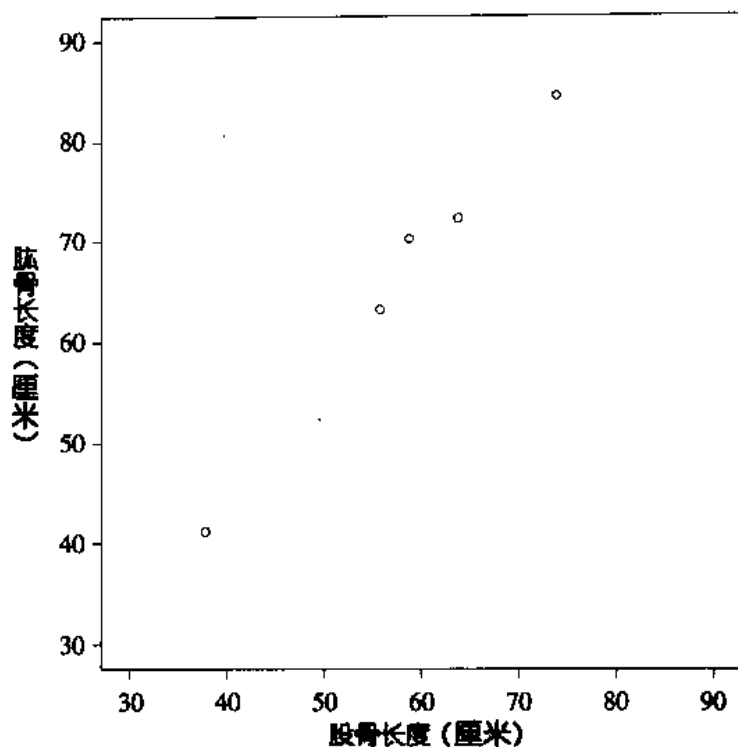


图 14.4 已灭绝的动物始祖鸟所遗留的五件化石标本中, 两种骨头(股骨与肱骨)长度散布图

相关系数

散布图呈现两个变量之间相关关系的方向、形式和强度。直线相关尤其重要, 因为直线是相当普遍的简单形态。当点的分布很接近直线时, 直线相关就很强, 而当点在直线附近散布很广时, 直线就弱。光用眼睛看, 不容易判断相关性有多强。图 14.5 的两个散布图画的是同样的一组数据, 只是右边的图坐标涵盖范围较大, 所以点变得比较靠近, 使得右边的图似乎显示出较强的直线相关。只要改一改散布图坐标轴上的刻度, 或者点和点之间的空白处的大小, 我们的眼睛就可能受骗。所以我们得遵照数据分析的一般策略, 除了图以外还要加上数值量度。相关系数(correlation)就是我们要用的量度。

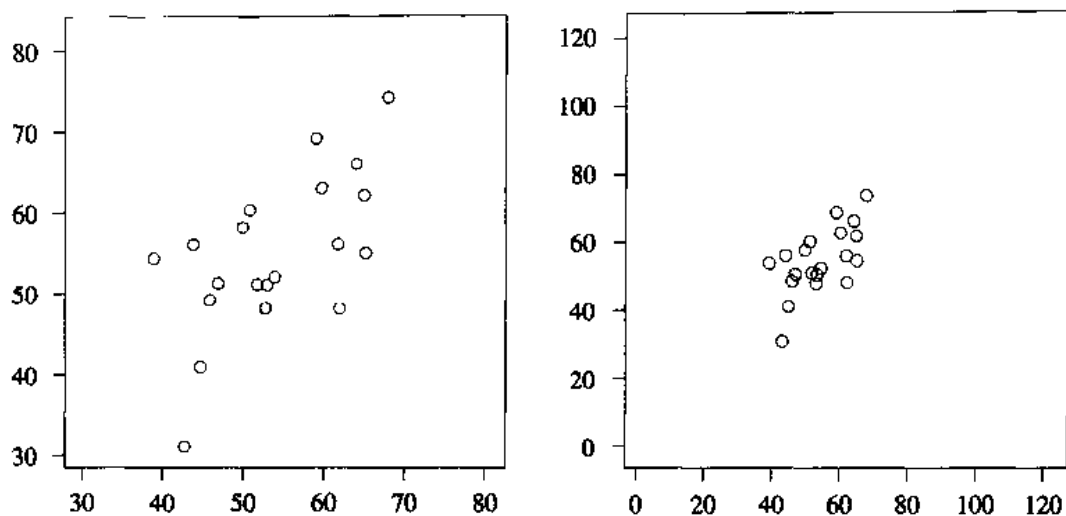


图 14.5 同一组数据的两个散布图。右图因为四周围空白较多,使得两个变量之间的相关性看起来比较强

• 相关系数

相关系数(correlation)描述两个数量变量(quantitative variable)之间直线相关的方向和强度。相关系数通常用符号 r 表示。

要计算相关系数得花点工夫。通常来说,你可以把 r 看成是按计算机的某个键或在软件中给某个指令就可以得到的数,而你只要了解它的性质和用处就可以了。但是如果知道 r 值是如何用数据算出来的,对于了解相关系数的性质和用处有很大帮助,所以我们还是举例告诉你 r 要怎么算。

例 3 计算相关系数

我们有 n 个个体的两种变量数据,分别叫做 x 和 y 。以例 2 的化石数据来说, x 是股骨长度, y 是肱骨长度,而我们有 $n=5$ 件化石的数据。

第一步:分别求出 x 和 y 的平均数和标准差。而化石数据,用计算机就得到如下数值:

股骨: $\bar{x} = 58.2$ 厘米, $s_x = 13.20$ 厘米

肱骨: $\bar{y} = 66.0$ 厘米, $s_y = 15.89$ 厘米



我们用 s_x 和 s_y 这样的符号, 是要提醒自己这里有两个不同的标准差, 一个对应变量 x 的值, 一个对应变量 y 的值。

第二步: 用从第一步得到的平均数和标准差, 求出每一个 x 值和每一个 y 值的标准计分。

x 值	标准计分 $(x - \bar{x}) / s_x$	y 值	标准计分 $(y - \bar{y}) / s_y$
38	$(38 - 58.2) / 13.20 = -1.530$	41	$(41 - 66.0) / 15.89 = -1.573$
56	$(56 - 58.2) / 13.20 = -0.167$	63	$(63 - 66.0) / 15.89 = -0.189$
59	$(59 - 58.2) / 13.20 = +0.061$	70	$(70 - 66.0) / 15.89 = +0.252$
64	$(64 - 58.2) / 13.20 = +0.439$	72	$(72 - 66.0) / 15.89 = +0.378$
74	$(74 - 58.2) / 13.20 = +1.197$	84	$(84 - 66.0) / 15.89 = +1.133$

第三步: 相关系数就是这些标准计分乘积的平均。就和算标准差时一样, 我们平均时的除数是 $n - 1$, 比个体的数目少 1。

$$\begin{aligned}
 r &= \frac{1}{4} [(1 - 1.530)(-1.573) + (-0.167)(-0.189) + (0.061)(0.252) \\
 &\quad + (0.439)(0.378) + (1.197)(1.133)] \\
 &= \frac{1}{4} (2.4067 + 0.0316 + 0.0154 + 0.1659 + 1.3562) \\
 &= \frac{3.9758}{4} = 0.994
 \end{aligned}$$

例 3 中的计算过程, 可以用以下的简化代数式表示

$$r = \frac{1}{n-1} \sum \left(\frac{x - \bar{x}}{s_x} \right) \left(\frac{y - \bar{y}}{s_y} \right)$$

其中 Σ 这个符号代表“全部加起来”。

了解相关系数的意义

比计算 r 值(这是机器的工作)更重要的是, 了解相关系数怎么度量相关性。以下是相关事实:



把数据画图之后，
要用用脑袋！

沃德(Abraham Wald, 1902—1950)和许多统计学家一样，在第二次世界大战时也处理了与战争相关的问题。他发明的一些统计方法，在战时被视为军事机密。以下是他提出的概念中较简单的一种。沃德被咨询飞机上什么部位应该加强钢板时，开始研究从战役中返航的军机上受敌军创伤的弹孔位置。他画了飞机的轮廓，并且标示出弹孔的位置。资料累积一段时间后，几乎把机身各部位都填满了。于是沃德提议，把剩下少数几个没有弹孔的部位补强。因为这些部位被击中的飞机都没有返航。

- 正的 r 值显示变量之间有正相关，负的 r 值显示出负相关。图 14.4 的散布图显示，股骨长度和肱骨长度之间有很强的正相关性。在其中三件化石中，两种骨头都比平均数要长，所以对 x 变量和 y 变量算出的标准计分都是正的。在另二件化石中，骨头长度都低于平均，所以 x 和 y 的标准计分都是负的。因此得到的乘积全是正的，使得 r 值为正。

- 相关系数 r 的值，永远在 -1 和 $+1$ 之间。 r 值若接近 0 ，代表很弱的直线相关。当 r 由 0 向 -1 或 $+1$ 趋近时，相关关系的强度会渐次增加。 r 值若接近 -1 或 $+1$ ，表示点的分布很接近一直线。而 $r = -1$ 及 $r = 1$ 这两个极端值的情况，只有散布图中的点全部落在同一条直线上时才会发生。

例 3 当中所得到的 $r = 0.994$ ，反映出图 14.4 的强烈正向直线形态。图 14.6 中的散布图说明了， r 怎样度量直线相关的方向和强度。你可以仔细研究一下这些图。请注意， r 的符号和每个图的倾斜方向一致，而且当图的形态愈来愈接近直线时， r 会愈来愈接近 -1 或 1 。

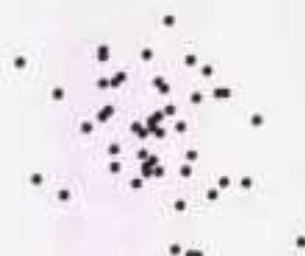
- 因为 r 是用观测值的标准计分算出来的，当我们分别或同时改变 x 、 y 的度量单位时， x 和 y 之间的相关系数并不会改变。例 3 当中的骨头长度如果用英寸而不用厘米来量，相关系数仍然会是 $r = 0.994$ 。

我们对于单一变量所做的各种描述性质用的量度，所使用的单位和原来的观测值相同。如果我们用厘米来量长度，则中位数、四分位数、平均数及标准差的单位全是厘米。不过两个变量间的相关系数可没有度量单位，它只是 -1 和 1 之间的一个数罢了。

- 相关系数不理睬解释变量和反应变量之间的差别。假如我们把 x 变量和 y 变量的名称对调，相关系数还是一样。
- 相关系数度量的只是两变量直线相关的强度。相关系数不能描述两变量间的曲线相关，不管这种相关关系有多强。
- 和平均数以及标准差一样，相关系数也会受少数异常观测值的严重影响。当散布图中出现异常点时，用起 r 来要特别小心。我们以图 14.7 来举例说明。假设我们把第一件化石的股骨长度从 38 厘米改成 60 厘米，这件化石的资料就不会和其他的大致成一直线，而会变成异常值了。而相关系数也从原来的 $r = 0.994$ 降到 $r = 0.640$ 。



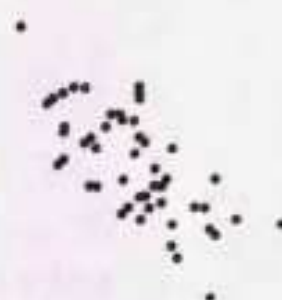
相关系数 $r = 0$



相关系数 $r = -0.3$



相关系数 $r = 0.5$



相关系数 $r = -0.7$



相关系数 $r = 0.9$



相关系数 $r = -0.99$

图 14.6 相关系数如何度量直线相关的强度。接近直线的形态，相关系数会趋近 1 或 -1

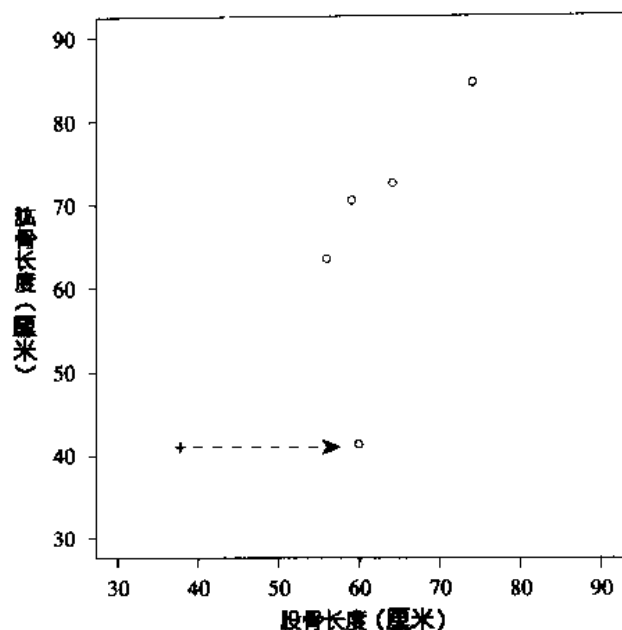


图 14.7 只移动一个点,就把相关系数从 $r=0.994$ 变成 $r=0.640$

变量之间的相关关系有许多种,要度量相关关系也有很多方法。虽然相关系数很常用,还是要记得它有其限制。相关系数只有对数量变量才有意义,也就是说我们可以讨论选民的性别和选民所属意的政党之间的相关关系,但是没法计算这两个变量间的相关系数。即使对于像骨头长度这样的数量变量来说,相关系数度量的也只是直线相关性。

而且还要记住,即使两个变量间有直线相关,相关系数也不是此两变量资料的完整描述。除了相关系数外,应该也列出 x 和 y 的平均数及标准差。因为计算相关系数的公式当中,有时会用到平均数和标准差,所以把它们随着相关系数一起列出也很恰当。

网络寻奇

要了解相关系数如何反映出散布图中点的分布形态,最好的方法是利用小程序,它可以让我们画出以及移动数据点,并立刻看到相关系数的变化。去《统计学的世界》原文版的网站, www.whfreeman.com/scc, 进入 Statistical Applets(统计小程序)找到 Correlation and Regression(相关与回归)的小程式。在你加入新的点或用鼠标移动某些点时,这个小程式会不断地重新计算相关系数。



本章重点摘要

大部分的统计研究都在探讨两个或多个变量之间的相关关系。**散布图**是展示两个数量变量之间相关关系的图形, 如果你要解释变量和反应变量, 应该把解释变量放在散布图的 x 轴(横轴)。

检视散布图的时候, 要找相关关系的方向、形式和强度以及可能有的**异常值**。如果方向很明确的话, 是正向的(图形从左到右往上斜)还是负向的(图往下斜)? 形式是直线还是曲线? 有没有观测值聚成一丛一丛的状况? 相关性很强(点所形成的形态很“扎实”)还是弱(点很散)?

相关系数 r 度量两个数量变量间直线相关的方向和强度。相关系数 r 是在 -1 — 1 之间的一个数, 它的符号显示出正相关还是负相关。当点愈聚集在一条直线的附近的时候, r 的值就愈接近 -1 或 1 , 而只有当散布图中的点全都落在同一条直线上时, r 的值才会是 -1 或 1 。



第 14 章 习题

14.1 我可以是哪些数字?

- (a) 相关系数 r 的所有可能值有哪些?
- (b) 标准差 s 的所有可能值有哪些?

14.2 度量蟋蟀。假设你为了一项生物作业, 度量了 12 只蟋蟀的长度(厘米)及重量(克)。

- (a) 说明为什么你预期长度和重量之间的相关系数会是正的。
- (b) 如果你用英寸当单位来量长度, 相关系数会怎么变?(1 英寸等于 2.54 厘米。)

14.3 IQ 和 GPA。图 14.8 中是美国中西部乡下一所学校全部共 78 位七年级生的学业平均(GPA, grade point average)对应 IQ 分数的散布图。

- (a) 以言语说明相关关系的整体形态。可以把 A、B、C 三点叫做异常值。

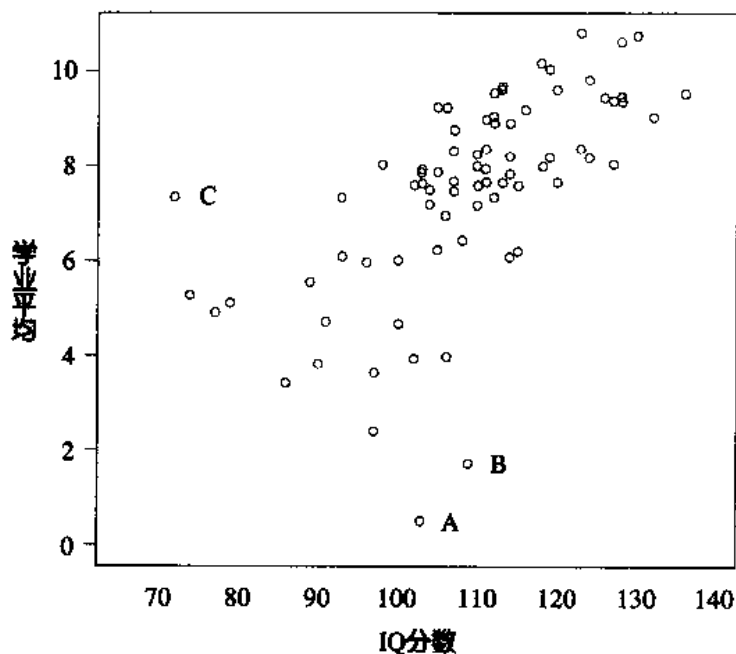


图 14.8 78 位七年级学生的学业平均和 IQ 测验分数, 对照习题 14.3



(b) 学生 A 的 IQ 和 GPA 各约为多少?

(c) 分别说明 A、B、C 三点的特别之处(比如“GPA 很低,但 IQ 分数却不差”之类)。

14.4 热狗的卡路里及含盐量。图 14.9 显示出 17 个品牌肉类热狗的卡路里数和钠含量。描述一下这组数据的整体形态,并指出点 A 有什么不寻常之处?

14.5 IQ 和 GPA。你认为图 14.8 中的数据之相关系数 r 应该是会接近 -1 ? 还是明显会是负值但不接近 -1 ? 或靠近 0 ? 明显为正但不接近 1 ? 还是会接近 1 ? 要说明你的答案。

14.6 热狗的卡路里及含盐量。你认为图 14.9 中的数据的相关系数 r 会接近 -1 ? 明显为负但不接近 -1 ? 靠近 0 ? 明显为正但不接近 1 ? 还是会接近 1 ? 要说明你的答案。

14.7 比较相关系数。图 14.8 和 14.9 比起来,哪一个的相关系数会比较接近 1 ? 说明你的答案。

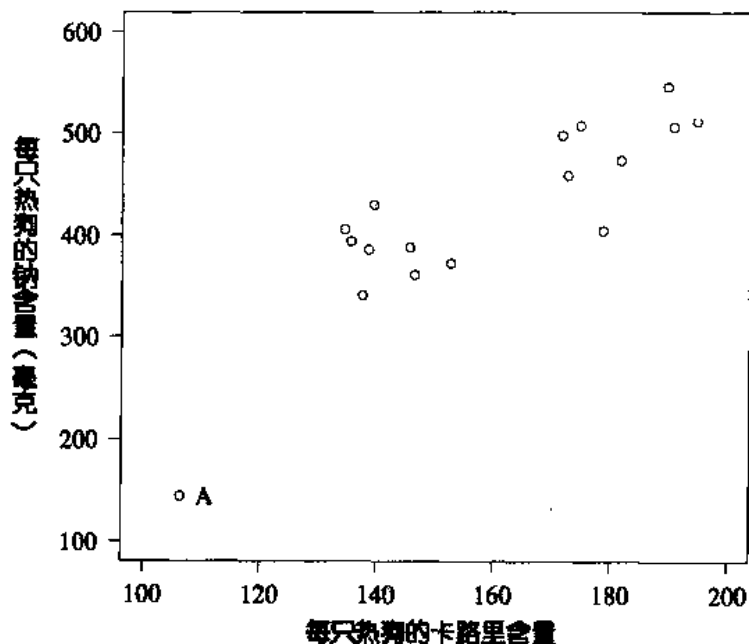


图 14.9 17 个品牌的肉类热狗的卡路里及钠含量, 对照习题 14.4



14.8 异常值和相关系数。图 14.8 里有 3 个异常值，分别用 A、B、C 标示。而图 14.9 里标示为 A 的点也是异常值。如果把异常值拿掉，则其中一个图的 r 值会增加，另一个 r 值会变小。请问哪个会增加，哪个会变小，为什么？

14.9 教授常游泳。穆尔教授为了摆脱中年形象，定期游泳且每次游泳两千码。以下是他游的时间(分钟)及游完后的脉搏(每分钟搏动次数)的 23 笔资料：

时间	34.12	35.72	34.72	34.05	34.13	35.72	36.17	35.57
脉搏	152	124	140	152	146	128	136	144
时间	35.37	35.57	35.43	36.05	34.85	34.70	34.75	33.93
脉搏	148	144	136	124	148	144	140	156
时间	34.60	34.00	34.35	35.62	35.68	35.28	35.97	
脉搏	136	148	148	132	124	132	139	

- (a) 画一个散布图。(解释变量是什么?)
- (b) 两个变量间的相关性应该是正还是负?请说明为何你做这样的结论。
- (c) 描述相关关系的形式和强度。

14.10 谁燃烧较多能量?在有关增重、节食及运动的研究中，新陈代谢率是很重要的指标，这指的是身体消耗能量的速率。表 14.1 列出了参与一项节食研究的 12 位女性和 7 位男性的瘦肉体重(LBM, lean body mass)及静止新陈代谢率。这儿的瘦肉体重以公斤为单位，是—个人把脂肪除外的体重。新陈代谢率是以每 24 小时燃烧的卡路里数来计算，这里的卡路里和用来描述食物含多少能量的卡路里是同样的。研究者相信瘦肉体重对新陈代谢率有重要影响。

- (a) 针对女性参与者的数据画个散布图。图中什么是解释变量?
- (b) 变量之间的相关性是正还是负?相关关系的形式是什么?强度又如何?
- (c) 现在把男性的数据也加进散布图里，用不一样的颜色或者符号。你在(b)中观察到的相关关系形态对男性适不适用?男性整体来说，和女性整体有何不同?



表 14.1 瘦肉体重及新陈代谢率

参与者	性别	男体重(千克)	代谢率(卡路里)	参与者	性别	体重(千克)	代谢率(卡路里)
1	男	62.0	1792	11	女	40.3	1189
2	男	62.9	1666	12	女	33.1	913
3	女	36.1	995	13	男	51.9	1460
4	女	54.6	1425	14	女	42.4	1124
5	女	48.5	1396	15	女	34.5	1052
6	女	42.0	1418	16	女	51.1	1347
7	男	47.4	1362	17	女	41.2	1204
8	女	50.6	1502	18	女	51.9	1867
9	女	42.0	1256	19	男	46.9	1439
10	男	48.7	1614				

14.11 婚姻。假设每位女性都和长她两岁的男性结婚。画一个 5 对夫妇年龄的散布图,以妻子的年龄为解释变量。你这组数据的相关系数 r 是多少?为什么?

14.12 拉开散布图 改变量度的单位可以大幅度改变散布图的外观。让我们回到例 2 的化石例子。

股骨	38	56	59	64	74
肱骨	41	63	70	72	84

这些数字是以厘米为单位量出来的。假设有位古怪的科学家用米为单位来量股骨,却用毫米为单位来量肱骨。结果得到以下数据:

股骨	0.38	0.56	0.59	0.64	0.74
肱骨	410	630	700	720	840

- (a) 画一个坐标,其中 x 轴的范围为 0—75, y 轴为 0—850,把原始数据画在这组坐标轴上。然后用不同颜色把新数据画在同一组坐标轴上。两个图看来应该很不一样。
- (b) 然而根据两组量度算出的相关系数应该完全相同。为什么你不用做任何计算就可以做这样的结论?



14.13 教授常游泳。习题 14.9 里有某位中年教授游泳 2 000 码所花时间和刚游完时脉搏的数据。

- (a) 用计算机找出相关系数 r 的值。用散布图来解释为何算出的 r 值是合理的。
- (b) 假设时间改用秒来度量。举例来说, 34.12 分钟就会变成 2 047 秒。 r 值会有什么改变?

14.14 谁燃烧较多能量? 表 14.1 有 12 位女性和 7 位男性的瘦肉体重和新陈代谢率。你在习题 14.6 里已画过这组数据的散布图。

- (a) 你觉得男性的相关系数和女性的相关系数会差不多还是颇有差距? 为什么?
- (b) 计算女性的相关系数和男性的相关系数。(用计算机。)

14.15 强相关性然而并无线性相关 汽车每加仑汽油跑的英里数在速度增加时先会上升再下降。假设这种相关关系相当规则, 如以下的速度(每小时英里数)和汽油里程(每加仑英里数)资料所示:

速度	20	30	40	50	60
汽油里程	24	28	30	28	24

画一个汽油里程对应速度的散布图。用计算机算一算, 速度和汽油里程之间的相关系数其实是 0。解释一下为什么虽然速度和汽油里程之间有很强的相关性, 但相关系数却是 0。

14.16 瘦肉体重和新陈代谢率。表 14.1 的瘦肉体重是以公斤为单位。1 千克等于 2.2 磅。假如我们把单位从千克改成磅, 平均瘦肉体重会有什么变化? 瘦肉体重和新陈代谢率之间的相关系数会有什么变化?

14.17 单位是什么? 你的数据包括若干参与者的年龄(以岁为单位)以及他们的反应时间(以秒为单位)。以下根据上述数据算出的各个统计量的单位是什么?

- (a) 参与者的平均年龄。
- (b) 参与者反应时间的标准差。
- (c) 年龄和反应时间之间的相关系数。



(d) 参与者的中位年龄。

14.18 教学与研究 一家大学报纸访问了一位心理学家,请他就学生对教授教学所做的评价发表意见。心理学家说:“证据显示,教授的研究成果多寡和教学评价好坏之间的相关系数接近0。”报纸却做了如下报道:“麦丹纽教授表示,研究好的教授多半教书差,而教书好的多半研究不行。”请说明为什么报纸这篇报道是错的。把心理学家所发表的意见用日常用语做个书面说明(不要用“相关系数”这种专有名词)。

14.19 关于相关系数的马虎叙述 以下每一条叙述中都有一个大错误。说明每一条的错误在哪里。

- (a) 美国就业者的性别和收入之间有很高的相关系数。
- (b) 我们发现在学生对教授的评价和其他同行对教授的评价之间,有很高的相关系数($r = 1.09$)。
- (c) 年龄和收入之间的相关系数,是 $r = 0.533$ 岁。

14.20 猜猜相关系数 从很大的样本算出的结果显示:

- (a) 父亲的身高和成年儿子身高之间的相关系数大约是_____。
- (b) 丈夫身高和其妻子身高之间的相关系数大约是_____。
- (c) 女性在4岁时的身高和18岁时身高之间的相关系数大约是_____。

答案是(没有依照顺序):

$$r = 0.2 \quad r = 0.5 \quad r = 0.8$$

把这些答案和题口配对,并说明为何做这样的选择。

14.21 猜猜相关系数 对以下几对变量,你会期望有明显的负相关、明显的正相关还是接近0的相关系数?

- (a) 二手车的车龄和车价。
- (b) 新车的重量和汽油里程(每加仑英里数)。
- (c) 成年男性的身高和体重。
- (d) 成年男性的身高和IQ。

14.22 分散投资。一家共同基金公司的新闻信上说:“多样化的投资组合应包含低相关的资产。”新闻信中还列出各种不同投资获利之



间的相关系数。举例来说，市政债券和高资本股票之间的相关系数是 0.50 而市政债券和低资本股票之间的相关系数是 0.21。

(a) 瑞秋对于市政债券做了大量投资。她想通过加入一个获利不会紧跟着她的债券走的投资项目来分散风险。为了达到这个目的，她应该选择高资本股票还是低资本股票？说明你的选择。

(b) 如果瑞秋想要找一种投资，是会在她的债券贬值时反而价值上升的，则应该要找怎样的相关系数？

14.23 带我去看球。大联盟棒球场卖的热狗和 16 盎司汽水的价格之间有何关系？表 14.2 里有部分资料。画一个可以用来显示汽水价格如何有助于解释热狗价格的散布图。描述一下你看到的相关关系。有没有异常值？

表 14.2 大联盟棒球比赛场地的热狗和汽水价格

队名	热狗	汽水	队名	热狗	汽水	队名	热狗	汽水
天使	2.50	1.75	巨人	2.75	2.17	游骑兵	2.00	2.00
太空人	2.00	2.00	印第安人	2.00	2.00	红袜	2.25	2.29
勇士	2.50	1.79	马林鱼	2.25	1.80	洛杉矶	2.25	2.25
酿酒人	2.00	2.00	大都会	2.50	2.50	皇家	1.75	1.99
红雀	3.50	2.00	教士	1.75	2.25	老虎	2.00	2.00
道奇	2.75	2.00	费城人	2.75	2.20	双城	2.50	2.22
博览会	1.75	2.00	海盗	1.75	1.75	白袜	2.00	2.00

14.24 降水量。图 14.10 里画的是美国各州有记录以来的年度最高降水量对应单日最高降水量的资料。代表阿拉斯加州(AK)，夏威夷(HI)和德州(TX)的点有特别之处标示出来。

(a) 阿拉斯加州的最高日降水量及年降水量大约各是多少？

(b) 相对于单日最高降水量来说，阿拉斯加和夏威夷有非常高的最高年降水量。把这两州当做异常值去掉。描述一下对其他州来说，相关关系有什么样的特征。如果知道某一州的最高日降水量，对于猜测该州的年度最高降水量帮助大不大？

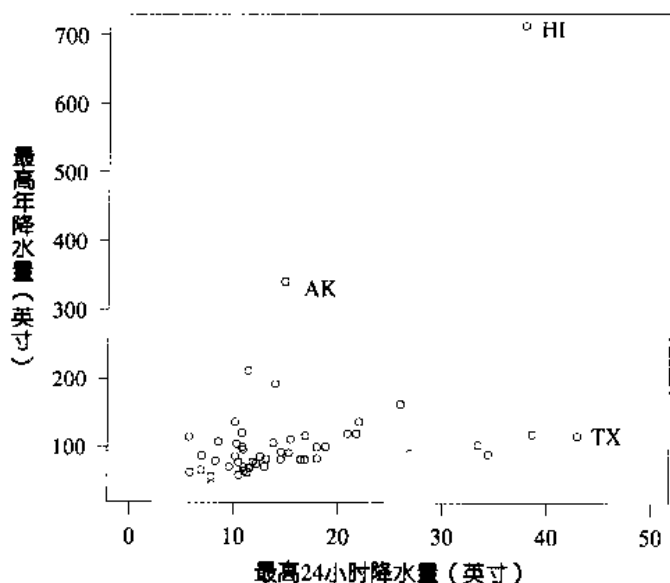


图 14.10 美国各州任一气象站有记录以来, 年度最高降水量对应于该州最高日降水量的散布图, 对照习题 14.24

14.25 要种多少玉米才算太多? 农夫应该在 1 英亩土地上种多少玉米, 才能得到最高的收获量? 要找出最佳的种植率, 应该做个实验: 在不同的农地上按不同比例种植玉米, 再度量收获量。以下是一个这类实验所得数据:

- 收获量和种植率哪个是解释变量?
- 画一个收获量和种植率的散布图。
- 描述相关关系的整体形态。是不是条直线? 有没有正相关, 或者负相关? 解释一下为什么增加每英亩的玉米株数会出现你图上所显示出的效应。

每英亩玉米株数	收获(每英亩蒲式耳数)			
12 000	150.1	113.0	118.4	142.6
16 000	160.9	120.7	135.2	149.8
20 000	165.3	130.1	139.6	149.9
24 000	134.7	138.4	156.1	
28 000	119.0	150.5		



14.26 为什么这么小?替下面这组数据画一个散布图。

x	1	2	3	4	10	10
y	1	3	3	5	1	11

用计算机算一算,相关系数应该是 0.5 左右。对这组数据中的大部分的点来说, x 和 y 之间有很强的直线相关,是什么因素导致相关系数只有 0.5 左右?

第 15 章

描述相关关系：回归、预测及因果关系

股票会涨还是会跌？

预测股市的走向可能让你发财。难怪有一大堆人和一大堆电脑都埋首在股市资料里寻找走势。

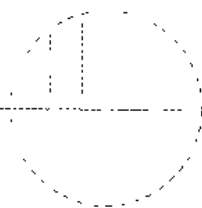
有些受欢迎的“名牌”有点匪夷所思。“超级杯指标”声明，每年一月份所举办的超级杯美式足球赛可以预测该年股市的表现。目前的美国国家足球联盟(NFL, National Football League)是由原来的 NFL 和美国足球联盟(AFL, American Football League)合并而成。超级杯指标声称，若原来属于 NFL 的球队赢了超级杯，该年股市就会上涨；若原属 AFL 的球队赢了，股市就会下跌。从 1967 年第一届超级杯起，到 1999 年为止的共 33 年期间，这个指标的预测有 28 次是正确



的。听起来好像很准。但是在那 33 年当中，股市只有 6 年是下跌的，所以只要每年都预测“上涨”，就会对 27 次。大部分的超级杯都是由原属 NFL 的球队得胜，一点儿也不稀奇：原属 NFL 的有 17 队，而 AFL 队只有 11 队，而且原来的 NFL 就是较有基础且较强的联盟。所以“NFL 赢了”的意义和“每年上涨”的意思差不多。20 世纪 90 年代的股市狂升，可一点儿也没有受 1998 和 1999 年 AFL 球队得胜的影响。

超级杯指标太没学问了。有些统计方法可以利用某些变量来预测另一变量的值，用到的资料不过只是几次上涨几次下跌而已。这些方法的源头叫做回归(regression)，我们将会在本章中讨论。而这些方法也牵涉甚广。某个网站收费预测股价，而其股价根据的是“人工神经网络、基因演算法、近邻模型及其他模型归类技巧，加上根据混沌、碎型及小波理论得到的股价的时间序列转换”。听起来好伟大，但是有用吗？给你点儿提示：如果这些方法真的可以预测股价，拥有它的人就不会卖给我们——他们会悄悄地用这些办法去发大财。

金融专家说，如果真的可以找出股价的未来走势，就会有许多的人试图凭借这个方法获利，结果很快就会把走势弄乱了。比如说，假如股市在圣帕特里克日(St. Patrick's Day, 3 月 17 日)和愚人节(4 月 1 日)之间通常都一路上涨的话，大家就会在 3 月 17 日前后买进股票，而在 4 月 1 日前后卖出。结果就是，因为大家在买，所以在那段期间刚开始时股价的确被往上推，但在要结束时大家都卖出会让股价下跌，就破坏了原来一路上涨的走势。这个逻辑应该比号称找到击败大盘妙法的声明要令人信服。在本章中我会常常提到：预测是件很微妙的事。



回归直线

如果散布图显示出两个数量变量之间的直线相关，我们会希望在散布图中画条直线，来对这个一般形态做概述。回归直线(regression line)就是对两个变量间的关系做概述，但条件是：其中一个变量可以用来解释或预测另一个变量。也就是说，回归描述的是一个解释变量和一个反应变量之间的相关关系。



• 回归直线

回归直线(regression line)是一条直线,描述当解释变量 x 的值改变时,反应变量 y 的值怎样跟着变。我们常用回归直线来预测:对于某一个给定的 x 值, y 值会是什么。

例 1 化石标本中的骨头

我们见到始祖鸟化石中,两种骨头的长度的相互关系很接近直线形式。图 15.1 中画出了 5 件标本的两种骨头长度。图里的直线对于整体形态做了简要的概述。

另外有件始祖鸟化石不完整,它的股骨有 50 厘米长,但是肱骨不见了。我们不能猜出肱骨有多长呢?连接肱骨长度和股骨长度的直线形式非常强,使得我们可以放心地用股骨长度来猜测肱骨长度。图 15.1 告诉我们怎么做:从股骨长度(50 厘米)开始,垂直往上直到和直线相交,然后从交点画水平线到代表肱骨长度的坐标轴,因此我们猜测长度大约是 56 厘米。如果代表这件化石的点恰好就在该条直线上的话,肱骨长度就会是这个数字。由于其他的点都离直线很接近,所以我们认为遗失了的肱骨所应属的点也会很接近直线。也就是说,我们觉得我们的猜测应该会很准。

例 2 总统选举

共和党的里根(Ronald Reagan)在 1980 和 1984 年两度当选美国总统。图 15.2 画出了各州投票给里根的民主党对手的百分比,这两位对手是:1980 年的卡特(Jimmy Carter)和 1984 年的蒙代尔(Walter Mondale)。图里显示出正向的直线相关。我们预期会有这种现象,因为有些州倾向于投民主党,而有些州倾向于投共和党。只有一个异常值:卡特总统的家乡佐治亚州,1980 年有 56% 投给民主党的卡特,但是 1984 年只有 40% 投给民主党。

我们可以用图 15.2 中所画的回归直线,根据 1980 年的投票结果,来预测某一



州1984年的投票情况。这个图里的点，比起图15.1里的化石骨头的点来说散布得离直线较远。度量直线相关强度的相关系数为 r ，在图15.1里 $r=0.994$ ，而在图15.2里则 $r=0.704$ 。从比较散开的点可以知道，要预测选举结果，一般来说准确度要比预测骨头长度来得差。

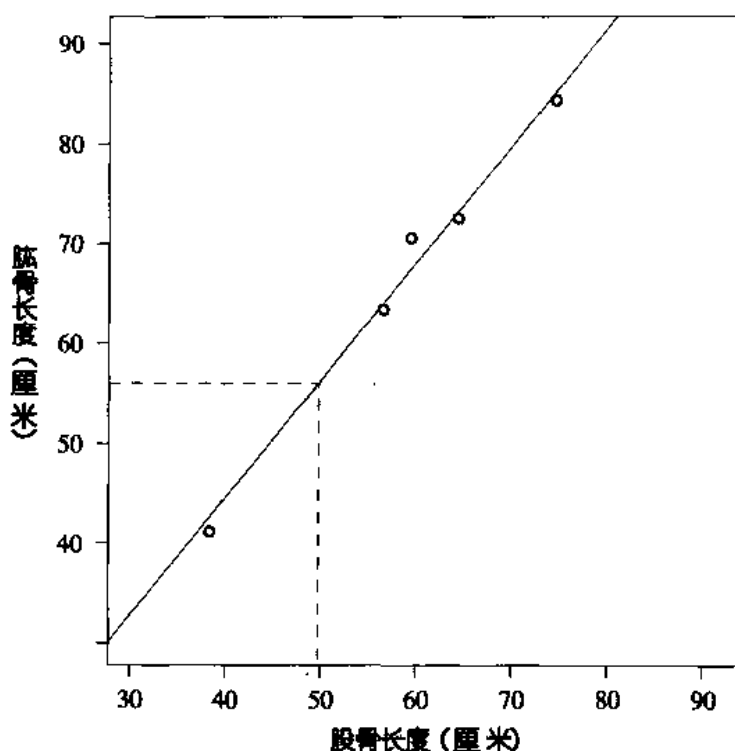


图15.1 用直线形式来做预测。数据是已绝种动物——始祖鸟的5件化石中两种骨头的长度

回归方程式

当散布图显示出像图15.1那么强的直线相关时，要用目测方式画一条接近所有点的直线是很容易的。然而对图15.2来说，不同的人用目测法，可能画出颇不一样的直线。因为我们是想用 x 来预测

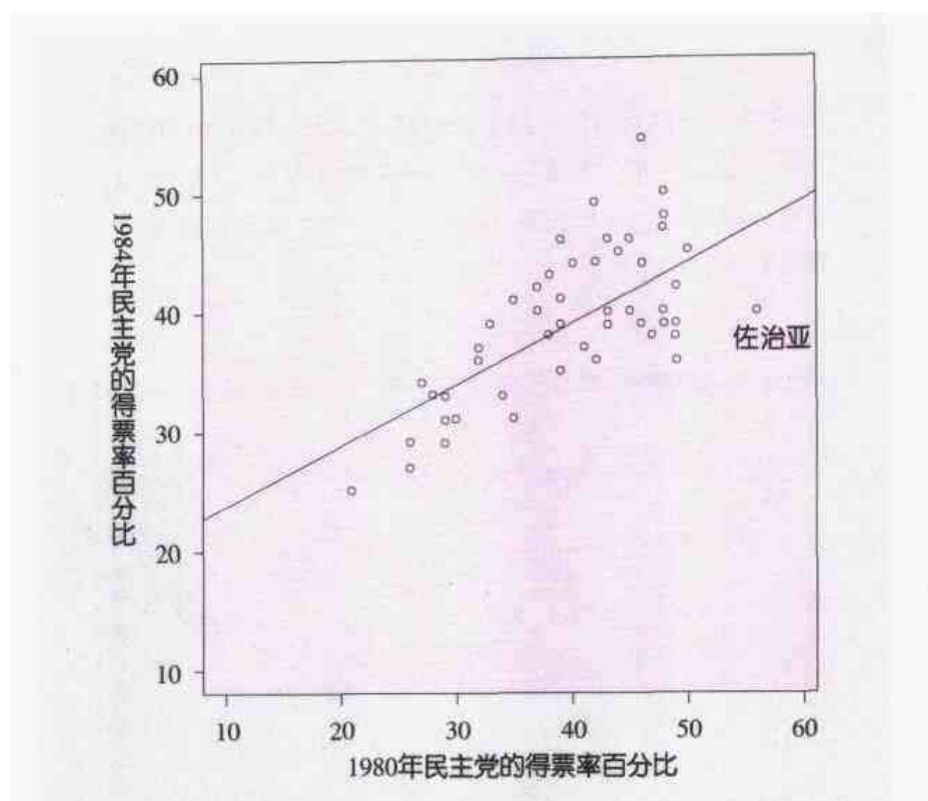


图 15.2 较弱的直线形态。图中数据是在里根两次竞选美国总统时，民主党在各州的得票率百分比

y ，所以我们想要的直线，是在垂直(vertical)方向(和 y 轴平行的方向)和点尽量接近。在用目测法画直线时，很难只去顾及点和直线的垂直距离。还不止这样，用目测法只能在图上得到直线，却得不到直线的方程式。我们需要有个办法，来根据数据找出垂直方向距点最近的直线方程式。有许多不同方法可以使垂直距离“越小越好”，而其中最常用的是**最小二乘法**(least-squares method)。

• 最小二乘法回归直线

y 对 x 的最小二乘法回归直线(least-squares regression line)，是使得所有数据点距直线的垂直距离平方和为最小的直线。

图 15.3 说明了最小二乘法的概念。这个图把图 15.1 的中间部分放大，焦点放在 3 个点上面。图里画出了这 3 个点距回归直线的垂直距离。要找最小二乘法回归直线必须用到所有的垂直距离(化石数据值全部 5 个点)，把每一个距离平方，然后移动直线，直到距离平方

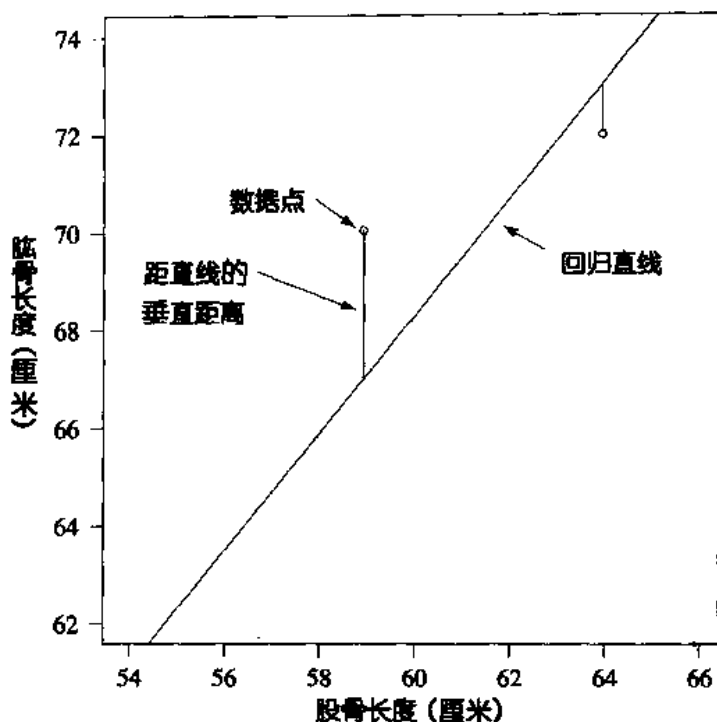


图 15.3 回归直线的目标是从 x 预测 y 。所以好的回归直线要让点到直线的垂直距离尽量地小

和的值达到最小为止。图 15.1 和 15.2 的散布图中所画的直线，就是最小二乘法回归直线。我们就不列出根据数据算出最小二乘法直线的公式了——这是电脑的工作。不过你应该会应用电脑所算出的方程式。

要写出这条直线方程式，还是像以前一样，让 x 代表解释变量，而 y 代表反应变量。方程式的形式如下：

$$y = a + bx$$

数字 b 是直线的斜率(slope)，就是 x 增加一个单位时 y 改变的量。数字 a 是截距(intercept)，是当 $x=0$ 时的 y 值。要利用这个方程式做预测，只要把你的 x 值代入方程式中，计算出 y 值即可。

朝平均数回归

“回归”(regress)这个词本来的意思是往回走。那为什么利用解释变量来预测反应变量的统计方法要用“回归”这个词呢？最先把回归方法用在生物及心理资料上的高尔顿爵士(Sir Francis Galton, 1822—1911)检视了诸如儿童身高对应其父母身高这类例子。他发现身高超过平均数的父母，通常孩子的身高也超过平均数，但是并没有父母那么高。高尔顿称这种现象为“朝平均数回归”，于是这个说法就被用在这种统计方法上了。

例 3 怎样应用回归方程式

在例 1 当中，我们在图 15.1 里用了“往上再往左”的



方法,预测了股骨长度为 50 厘米的化石,所应对应的肱骨长度。最小二乘法直线的方程式是:

$$\text{肱骨长度} = -3.66 + (1.197 \times \text{股骨长度})$$

这条直线的斜率是 $b = 1.197$ 。这代表对于这些化石来说,股骨长度每增加 1 厘米,肱骨长度就会增加 1.197 厘米。回归直线的斜率对于了解数据通常很重要。斜率是变化率(rate of change),即是指当 x 增加 1 时,我们预测的 y 所改变的量。

最小二乘法直线的截距是 $a = -3.66$ 。这是当 $x = 0$ 时,我们预测的 y 值。虽然要画出直线需要知道截距,但是只有当 x 值实际上有可能靠近 0 时,截距才有统计上的意义。这儿的股骨长度不可能是 0,所以截距没有统计上的意义。

要用方程式来做预测,只要把 x 值代入式子算出 y 即可。对应 50 厘米长的股骨化石,肱骨长度的预测值是:

$$\begin{aligned}\text{肱骨长度} &= -3.66 + (1.197)(50) \\ &= 56.2 \text{ 厘米}\end{aligned}$$

要在散布图上画出这条直线的话,用两个不同的 x 值分别预测出 y 值。这样就得到两个点。把这两个点画在图上,再连接两点画条直线即可。

了解预测的意义

电脑使得预测很容易而且全自动,即使对大笔的资料仍然一样。任何可以用全自动方式处理的事,处理时通常是不经过思考程序的。比如说,即使资料之间有曲线相关,回归软件仍然“乐于”匹配(fitting)一条直线。还有,电脑也不会自己决定谁是解释变量、谁是反应变量。这点很重要,因为同一组数据如果解释变量不同,会得出



两条不一样的直线。

在实际应用时，我们常常用多个解释变量来预测一个反应变量。大学在处理入学申请时，可能会用 SAT 的数学及语言分数，再加上高中英文、数学及科学成绩（共 5 个解释变量）来预测学生的大一表现。虽然细节很复杂，但是所有用来预测反应变量的统计方法，都和最小二乘法回归直线有一些共同的基本性质。

- 预测根据的是对数据匹配的某个“模型”（model）。在图 15.1 和 15.2 里，我们的模型就是我们穿过散布图中的点所画的一条直线。其他的预测方法使用较复杂的模型。
- 模型匹配得离数据点很接近的，预测结果最好。比如图 15.1 和图 15.2，前者所含的点距直线很接近，后者则否，所以图 15.1 的预测比较可靠。当变量多的时候，形态就不容易看出来，而且只要数据没有呈现出很强的形态，预测就可能很不准。
- 预测超出现有数据的范围是很靠不住的。假设你手上有 3—8 岁孩童的生长资料，你发现年龄 x 和身高 y 之间有很强的直线相关。如果你对这些数据匹配一条回归直线，然后用它来预测 25 岁时的身高，你的预测会是：这个孩子 25 岁时会有 8 英尺高。人长高到某个阶段会慢下来，然后会完全停止，所以把直线一直延长到成人的年龄是很笨的做法，没有人在预测身高时会犯这种错。但是几乎所有的经济预测都在试图告诉我们下一季或下一年会发生什么事，难怪经济预测常常错。

计算选票的人有没有作弊？

在宾夕法尼亚州的选举中，根据投票机的计数，共和党的马克斯领先民主党的史汀森。但是在控制选举委员会的民主党人计算了缺席投票的选票后，变成史汀森领先。事情闹上了法庭。法庭召唤一位统计学家，他用过去的选举资料造出回归直线，再根据投票机结果，预测缺席选票的计数。根据马克斯在投票机所领先的 564 票，可以预测他应该比史汀森多得 133 张缺席选票。实际上投票机算出来的，即是史汀森比马克斯多得了 1 025 张缺席选票。你想计算选票的人有没有作弊？

例 4 预测财政盈余

美国国会预算委员会必须每年提出报告，预测下五年的联邦预算以及盈余或者赤字。这些预测和未来的经济趋势（未知）有关，也和国会对税收和开支的决定（也未知）有关。而即使目前政策都不变，要预测预算的状况都会非常不准确。比如说，1996 年做的预测，就有超过 3 000 亿美元的出入。连 1997 年对次年所作的预测，都相差了 1 020 亿。就如参议员德克森（Everett Dirksen）曾说的，“这里差 10 亿，那里差 10 亿，很快就要真的出问题了”。1999 年预算



委员会预测接下来的 10 年会有 9 960 亿美元的盈余(不考虑社会保险)。政客们已经在讨论怎么用这些钱,但其他人都不相信这项预测。

相关系数及回归

相关系数度量直线相关的方向和强度,而回归画出一条直线来描述这个相关关系。相关系数和回归是密切相关的,即使回归需要选择解释变量而相关系数不需要。

相关系数和回归都会受异常值的严重影响。如果你的散布图有明显的异常值就要小心。图 15.4 里画的是美国各州年降水量最高纪录对应 24 小时降水量最高纪录。夏威夷是位于图的高处的异常值,它

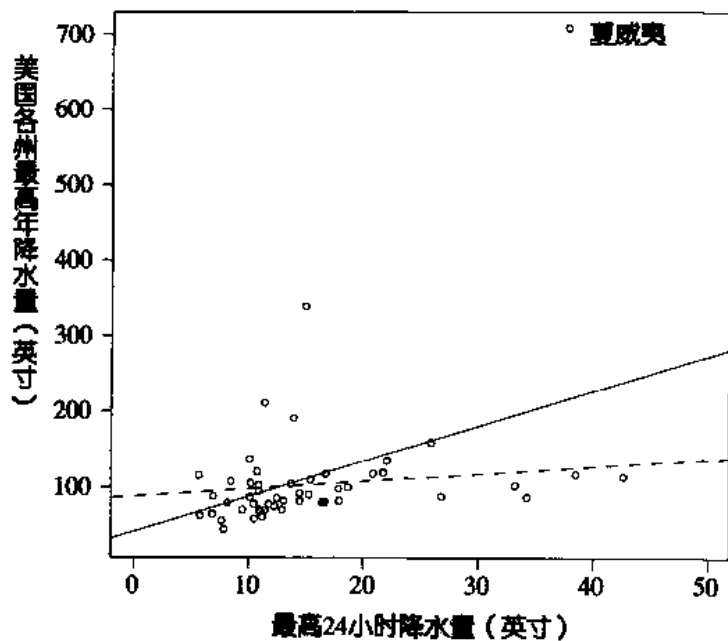


图 15.4 最小二乘法回归直线受异常值的严重影响。实线是根据全部 50 个数据点画的。虚线则去除了夏威夷



于1982年曾在库库伊(Kukui)记录到704.83英寸的年雨量。图15.4里所有50州的相关系数是0.408。如果把夏威夷去掉，相关系数会降到 $r=0.195$ 。图里面的实线，是从24小时纪录所预测年纪录的最小二乘法回归直线。如果不计入夏威夷，最小二乘法回归直线就会往下落到虚线的位置。这条直线(虚线)差不多接近水平，也就是说一旦我们决定忽略掉夏威夷，年降水量纪录和24小时降水量纪录之间就没有多大关系了。

回归直线的预测功能，视相关关系强度而定。也就是说，一条回归直线有多人用处，和变量之间的相关系数密切相关。事实上这个关系就是用相关系数的平方来度量的。

回归里的 r^2

相关系数的平方， r^2 ，是 y 值的变异当中，可以用 y 对 x 的最小二乘法回归来解释的部分所占之比例。

这背后的概念是说，当 y 和 x 有直线关系时， y 的变异中的一部分，可以解释为当 x 改变时，把 y 也拉着一起改变。

例5 r^2 怎么用

再来看看图15.1。这5件化石的肱骨长度变异很大，最低的是41厘米而最高的是84厘米。从散布图可以看出，我们只要看看肱骨长度以及回归直线，就几乎可以解释所有的变异了。当股骨长度增加时，它会沿着直线把肱骨长度也一块儿拉长了。除此之外，肱骨长度的变异就没剩下多少，剩下的这些变异从图上来看，就是指点和直线还是有些距离，而未直接落在直线上。因这组数据的 $r=0.994$ ，所以 $r^2=(0.994)^2=0.988$ 。因此由股骨拉着肱骨沿着直线上升时肱骨产生的变异，可以解释肱骨长度变异中的98.8%。点在直线两侧的散布，只能说明剩下1.2%的变异。剩下的散布很小，表示预测会很准。

再对照图15.2的投票资料。1980年和1984年的民主党得票率之间还是有直线相互关系，但是点在直线两侧也散布得比较开。这里的 $r=0.704$ ，而 $r^2=0.496$ ，



我们观察到的 1984 年民主党得票率的变异，大约只有一半可以用直线形态来解释。把 1980 年民主党得票率 45% 的州和得票率 30% 的州做个比较，你还是会猜到前者在 1984 年的民主党得票率较高。但是在 1980 年两党都有相同得票率的各州，1984 年的得票率有不小的变异，这就是 1984 年各州变异的另外一半。造成这部分变异的是其他原因，诸如两次选举的主要议题有别，以及里根总统的两位民主党对手来自不同地区等。

通常在报告计算出来的回归直线时，也会同时提出 r^2 的值，当做回归直线解释反应变量有多成功的一种指标。当你看到一个相关系数的时候，可以把它平方，会更容易感受相关的强度。完美的相关系数 ($r = -1$ 或 $r = 1$) 代表所有的点全落在一直线上，此时 $r^2 = 1$ ，而一个变量的所有变异，都可以用它和另一变量的直线相关关系来说明。若 $r = -0.7$ 或 $r = 0.7$ ，则 $r^2 = 0.49$ ，而差不多一半的变异可以用直线相关来解释。以 r^2 的值当标准的话，相关系数 ± 0.7 差不多在 0 和 ± 1 的中间。

因果问题

抽烟和肺癌死亡率之间有很强的相关性。是不是抽烟导致肺癌呢？在一个国家里，容不容易取得手枪和该国枪杀事件的比例之间也有很强的相关性。容易取得手枪是否导致更多谋杀案？香烟包装上已明白写着吸烟导致肺癌。有更多的人拥有手枪是否导致更多谋杀却引起热烈的辩论。为什么香烟和肺癌的证据优于手枪和杀人的证据呢？

我们已经知道统计证据中与因果关系有关的三大事实。

• 统计及因果

1. 即使两个变数间有很强的相关性，也并不一定代表改变其中一个变量的值会导致另一个变量的改变。
2. 两个变量之间的相关性，常常受其他潜藏在背景中的一些变量影响。
3. 建立因果关系最好的证据，来自随机化比较实验。



例6 看电视会延年益寿？

统计一下世界各国平均每人拥有电视机数 x 及人民预期寿命 y 。你会找到很高的正相关：有很多电视机的国家，人民预期寿命比较长。

因果关系的基本意义是，只要改变 x 的值，就可以使 y 的值改变。我们能不能运一堆电视机到博茨瓦纳，来延长那儿人民的寿命呢？当然不行。富国的电视机比穷国多，而富国的人民预期寿命也比较长，但这是因为他们有较好的营养、干净的水以及较好的医疗资源。电视机和寿命长短之间并没有因果关系。

例6 说明了我们三大事实的头两项。这一类的相关有人叫它做“胡说相关”：相关是事实，胡说的部分是“改变其中一个变量会导致另一个变量改变”的结论。像例6中的国家财富这种潜在变量(lurking variable)会同时影响 x 和 y 的值，造成 x 和 y 之间的高相关，即使 x 和 y 彼此之间并没有什么直接关系。我们可以把这个叫做共同反应(common response)：解释变量和反应变量都对某个潜在变量产生反应。



“依照第三世界的新脱贫计划，援助组织今天开始送出 10 000 台电视机。”



枪械管制和犯罪

严格管制枪械，尤其是手枪，是不是能减少犯罪？对许多人来说，答案一定为“是”。美国所有的谋杀案中，超过一半的凶器是手枪。美国的谋杀率（以每 100 000 人口计）是加拿大的 1.7 倍，而以手枪为凶器的谋杀率，更是高过加拿大 15 倍。显然手枪助长了坏事发生。芝加哥大学的经济学家洛特(John Lott)做了一项大规模的统计研究，用了 1977—1994 年的 18 年间美国所有 3 054 个县的资料。洛特发现，当许多州放松枪械管制，允许成年人携带枪支之后，犯罪率降低了。他的论点是，拥有枪支让老百姓可以自卫，也让罪犯有所畏惧，所以犯罪率降低了。

洛特用了回归方法来决定犯罪和许多解释变量之间的关联，并且经由对其他解释变量的调整，将允许携带隐藏着的枪支的效应分离出来。之后引发的争论非常热烈，到现在还在进行。人们对于枪械管制的反应很强烈，而大部分人对洛特研究的反应，是根据他们喜不喜欢他的结论而定的。支持可以拥有枪支的人把洛特形容成摩西，终于将真相呈现在世人面前；反对者却坚认他既不对、又邪恶。

洛特到底对不对？我不知道。他的研究比起大部分以前得出支持管制枪械的研究，

要更复杂精密。然而大型的观测研究有许多可能的弱点，尤其当该研究是在寻找随时间出现的趋势时。18 年中发生了许多事情，而有些并没有被纳入洛特的模型来考虑——举例来说，警察对于查缉非法枪械比以前积极。好的数据并不容易取得——比如说，合法拥有枪械的人数低于政府发出的准许拥有枪支的执照数，且实际人数很难精确估计。因此必须做非常仔细的研究，才能对洛特作的统计做出合理的评估。

质疑洛特之发现的最好理由，包括对于观测研究弱点的了解，以及事实上也有设计更精良的统计研究，做出应该减少枪支流通的结论。在哥伦比亚国内数个城市暂时性的禁止携枪（此次行动经过大力宣传，并由警察设据点执行搜索）和此前未执行禁止携枪时相比，的确降低了谋杀率。堪萨斯市枪支实验比较了两个犯罪率高的地区。在其中一区，警察在汽车停下时或者有人违规时就搜车并没收枪支。和枪支有关在犯罪的处理区降了一半，而在控制区（即对照组）却没变。因此似乎有理由相信，降低非法持枪可以减少使用枪支的罪行。当然洛特讨论的是合法持枪。于是就像许多因果问题一样，这个问题还没解决。



例7 肥妈妈和胖女儿

是什么造成儿童的肥胖？父母遗传、吃的太多、活动太少和看太多电视都会被当做解释变量。

一项对92个墨西哥裔美国女孩子做的研究，有典型的结果。研究当中度量了女孩子和她们妈妈的体脂肪健康指数(BMI, body mass index)，这是体重相对于高度的一种量度。BMI过高的人被认为过重或者肥胖。他们还度量了看电视的时间长短、体力活动的时间长短和数种食物的摄取量。所得到的结果是：女孩子的BMI和体力活动只有弱相关($r = -0.18$)，和食物及电视也是弱相关。最强的相关($r = 0.506$)出现在女儿的BMI和妈妈的BMI之间。

一个人的体型有部分决定于遗传。女儿从妈妈处继承了一半的基因，因此妈妈和女儿的BMI之间，有直接的因果关系。当然这个因果关系不是百分之百。妈妈的BMI只解释了女儿BMI变异的25.6% (这是 r^2)。还有其他因素也会影响BMI，其中有的因素在此研究中有的考虑到了，而有的则没有。即使有直接的因果关系，也极少可以完全解释两个变量之间的相关。

我们可不可以用例7里的 r 或 r^2 来说明，女儿的BMI有多少成分是由遗传造成的呢？答案是不肯定的。记住有交叉这回事。很可能过重的妈妈对于女儿来说，也是很少运动、饮食习惯不良及看很多电视的一个“榜样”。她们的女儿或多或少也养成同样的习惯，所以遗传的影响和女孩子周围环境的影响就混在一起了。我们没法判断，妈妈和女儿BMI之间的相关系数，有多大部分应该归因于遗传。

图15.5用简图显示：变量之间的相关，有可能用很多不同的关系来解释。虚线代表 x 和 y 变量之间已观测到的相关关系。有些相关可以用变量之间直接的因果关系来解释。图15.5的第一个图中，用从 x 到 y 的箭头表示“ x 导致 y ”。第二个图说明了“共同反应”。在 x 和 y 变量间观测到的相箭，可用潜在变量 z 解释，即 x 和 y 都会因为 z 的改变而改变。这种共同反应会制造出一种相关关系，即使 x 和 y 之间也许并没有直接的因果关系。图15.5中的第三

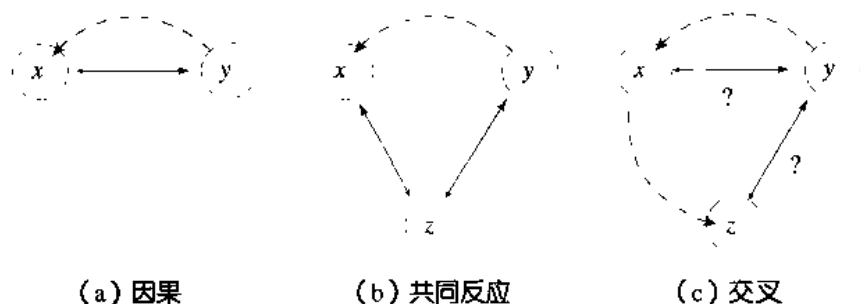


图 15.5 被观测所得到的相关性的部分解释。虚线表示相关，箭头显示因果关系。 x 是解释变量， y 是反应变量， z 是潜在变量

个图说明了交叉。解释变量 x 及潜在变量 z 可能同时影响反应变量 y 。因为变量 x 和 z 有关系，我们没法把 x 对 y 的影响和 z 对 y 的影响分离。我们没法说出 x 对 y 的直接影响有多大。事实上， x 对 y 是不是有影响，可能都很难说。

共同反应和交叉二者，都牵涉到潜在变量 z 对反应变量 y 的影响。我们不多谈这两类相关关系有什么差别，只要记住在考虑变量之间的相关关系时，要遵从“小心潜在变量”的忠告。以下是另一个共同反应的例子，这回我们想要做的是预测。

例 8 SAT 分数和大学成绩

高中时在 SAT 测验得了高分，当然并不能造成在大学里的好成绩，两者间的中等相关(r^2 大约是 27%)，无疑地可以用“对于诸如学习能力、读书习惯及不喝酒这类潜在变量的共同反应”来解释。

SAT 分数是否可以用来部分预测大学生的表现，和因果并没关系。我们只需要相信，过去几年间我们看到的 SAT 分数和大学成绩的相关性，对于今年毕业的高中生来说仍没改变。再想想我们的化石例子，用股骨长度可以很准确的预测肱骨长度。两种骨头间的强相关，可以解释为对于我们正在检视的化石所属始祖鸟的年龄和体型大小的共同反应。做预测不需要有因果关系。

对这些例子的讨论，又引出了关于因果的两大事实：



统计及因果的再补充

4. 在二个变量间观察到的相互关系，可能来自于**直接因果关系**(direct causation)、**共同反应**(common response)或是交叉。有可能其中两种因素或全部三种因素都同时存在。
5. 观察到的相互关系不管是不是因果，都可以拿来作预测，只要从以前的数据找出来的形态仍然适用。

因果证据

虽然有许多困难，但在没做实验的状况下，有时仍然可能得到很强的证据，指向因果关系。在非实验的证据当中，吸烟会导致肺癌的证据已经是强到不能再强了。

医师早就观察到，大部分肺癌病患是吸烟者。将吸烟者和“类似的”不吸烟者利用观测研究做比较的结果，显示吸烟与死于肺癌之间有强的相关性。这个相关性可不可能用该研究无法度量的潜在变量来解释呢？比如说，可不可能有某种遗传因子，会使人既容易对尼古丁上瘾，又容易得肺癌？这样的话，即使吸烟对肺癌没有直接影响，吸烟和肺癌还是会有正相关关系，这些反对意见是怎么克服的呢？

我们来回答更一般的问题：当我们不能做实验时，能够确立因果关系的标准在哪里？

- **相关性很强。**吸烟和肺癌之间的相关性很强。
- **相关关系有一致性。**在许多国家对不同的人所做的多项研究，都把吸烟和肺癌联系起来。这就降低了用“只和某群人或某研究有关的潜在变量”来解释相关性的可能。
- **较高剂量和较强反应相关。**每天吸烟较多的人或者吸烟历史较久的人更常得肺癌，而戒烟的人风险就降低。
- **被怀疑的原因在时间上超前结果。**肺癌是在吸烟多年后显现的。死于肺癌的男性人数在吸烟者人口普遍之后上升，时间的差距约有30年。男性死于肺癌的人数比死于其他任何癌症的都多。在女性开始吸烟之前，很少有女性得肺癌。女性肺癌患者的人数也随着吸烟人口的增加而增加，中间的差距也是30年，现在肺癌已超过



乳癌，成为女性癌病中的头号杀手。

- **被怀疑的原因是可信的。**动物实验的结果显示，吸烟产生的焦油的确会致癌。

医界当局毫不犹豫地宣称吸烟导致肺癌。美国公共卫生署署长早就声称吸烟是“在美国最大可避免的致死及致残原因”。因果证据是压倒性的——但是比不上用设计完美的实验所得到的证据强。

网络寻奇

想要感受一下怎样在散布图的点之间画出一条回归直线，最好的方法是利用一种小程序，它让你画一些数据点，并可以移动某些点而立即看到回归直线如何随着点的改变而移动。访问本书原文版的网站，www.whfreeman.com/sec，点选“Correlation and Regression”小程序。只要在“Show least-squares line”选项按一下左键，就可以看到回归直线。



本章重点摘要

回归是统计方法的名称，这类方法替数据匹配模型，以便根据一个或多个解释变量来预测反应变量的值。最简单的回归是在散布图上配备一条直线，用来由 x 预测 y 。配备直线最常用的方法是**最小二乘法**，用这个方法找到的直线，使数据点距直线垂直距离的平方和为最小。

最小二乘法回归与相关系数的关系密切，明确一点说，**相关系数的平方 r^2** 告诉我们，反应变量 y 的变异中，有多少比例可以用 y 和 x 的直线相关来解释。一般来说，统计预测准不准，是看数据之间有没有很强的形态而定。对超过数据范围外的部分进行预测是很冒险的，因为范围内的形态未必延伸到范围外。

两个变量间有强的相关性，不见得代表改变其中一个变量的值，会导致另一变量值的改变。潜在变量可以通过**共同反应或交叉**制造出关联。如果没办法做实验，通常很难得到足以令人信服的证据来证明因果关系。



第 15 章 习题

15.1 肥妈妈和胖女儿。例 7 中的研究发现，女孩子的 BMI 和一天中体力活动时间长短之间的相关系数是 $r = -0.18$ 。为什么我们会预期这个相关系数是负的？研究中女孩子 BMI 的变异，有多少百分比可以用和活动时间长短的直线相关关系来解释？

15.2 各州 SAT 分数。图 14.2 画出了美国各州高中应届毕业生在 SAT 数学部分的平均分数，对应参加测验学生的百分比。除了两个群外，图中还显示出整体的粗略直线形态。用参加测验的百分比来预测 SAT 数学分数的最小二乘法回归直线是：

$$\text{SAT 分数} = 574.6 - (1.102 \times \text{参加测验百分比})$$

(a) 斜率 $b = -1.102$ 告诉了我们有关这两个变量之间相关的什么信息？

(b) 在纽约州有 76% 的高中应届毕业生参加了 SAT 测验。请预测他们的平均数学分数。（纽约的实际平均分数是 503。）

15.3 IQ 和 GPA。图 14.8 画出了 78 位七年级学生的学业平均对应 IQ 测验分数。这些点虽有粗略的直线形态，然而也有不少的散布。两个变量间的相关系数是 $r = 0.634$ 。78 位学生 GPA 之间的变异，有多少百分比可以用 GPA 和 IQ 之间的直线相关来说明？有多少百分比可以用有类似 IQ 分数的学生之间的 GPA 差异来解释？

15.4 各州 SAT 分数。各州平均 SAT 数学分数和高中生参加 SAT 测验的百分比之间的相关系数是 $r = -0.897$ 。

(a) 相关系数是负的。这告诉了我们什么？

(b) 用参加测验的学生所占百分比来预测平均分数，效果会如何？（用 r^2 来回答。）

15.5 IQ 和 GPA。根据画在图 14.8 里的 78 位学生资料，做出可从 IQ 分数预测 GPA 的最小二乘法回归直线为：

$$\text{GPA} = -3.56 + (0.101 \times \text{IQ})$$

说明斜率 $b = 0.101$ 代表的意义，并预测 IQ 为 115 的学生，GPA 应是多少。



15.6 教授常游泳。以下是穆尔教授游2 000码所花的时间(分钟)以及游完泳之后的脉搏次数(每分钟)，总共23笔资料：

时间	34.12	35.72	34.72	34.05	34.13	35.72	36.17	35.57
脉搏	152	124	140	152	146	128	136	144
时间	35.37	35.57	35.43	36.05	34.85	34.70	34.75	33.93
脉搏	148	144	136	124	148	144	140	156
时间	34.60	34.00	34.35	35.62	35.68	35.28	35.97	
脉搏	136	148	148	132	124	132	139	

你在习题14.9已经画了这组数据的散布图。最小二乘法回归直线是：

$$\text{脉搏} = 479.9 - (9.695 \times \text{时间})$$

隔天他游了34.30分钟，请预测教授的脉搏。他的脉搏事实上是152。你的预测很准吗？

15.7 葡萄酒和心脏病。适量饮用葡萄酒可以预防心脏病。我们来看看一些国家的资料。表15.1中是19个发达国家一年的葡萄酒消耗

表15.1 葡萄酒消耗量及心脏病死亡人数

国家	从葡萄酒得到的酒精*	心脏病死亡率*
澳大利亚	2.5	211
奥地利	3.9	167
比利时/卢森堡	2.9	131
加拿大	2.4	191
丹麦	2.9	220
芬兰	0.8	297
法国	9.1	71
冰岛	0.8	211
爱尔兰	0.7	300
意大利	7.9	107
荷兰	1.8	167
新西兰	1.9	266
挪威	0.8	227
西班牙	6.5	86
瑞典	1.6	207
瑞士	5.8	115
英国	1.3	285
美国	1.2	199
德国	2.7	172

* 每人从喝葡萄酒所摄取酒精的升数

+ 每十万人死亡人数



量(平均每人喝葡萄酒摄取酒精的升数)以及一年中因心脏病死亡的人数(每 10 万人死亡人数)。

- (a) 画一个散布图来显示,一国的葡萄酒消耗量如何有助于解释心脏病的死亡率。
- (b) 说明相关关系的方向、形式及强度。
- (c) 这里两个变量间的相关系数是 $r = -0.843$ 。为什么这个数字符合你在(b)部分的描述?

15.8 山狸与甲虫 生态学家有时会在我们的环境中发现一些颇奇怪的关系。有一项研究似乎显示出,山狸对甲虫有益。研究者规划了 23 块圆形的地,每一块的直径都是 4 米,这些地都位于山狸会啃咬三角叶杨木的地区。在每一块地上,他们统计了被山狸啃断了的树桩个数以及甲虫幼虫的群数。数据如下:

树桩	2	2	1	3	3	4	3	1	2	5	1	3
幼虫群	10	30	12	24	36	40	43	11	27	56	18	40

树桩	2	1	2	2	1	1	4	1	2	1	4
幼虫群	25	8	21	14	16	5	54	9	13	14	50

- (a) 画一个散布图来显示出山狸啃咬的树桩个数如何影响甲虫幼虫的群数。从你的图里看出什么?(生态学家认为,树桩长出的新芽比其他三角叶杨木长出的新芽要嫩,所以甲虫比较喜欢。)
- (b) 最小二乘法回归直线是:

$$\text{幼虫群数} = -126.8 + (11.894 \times \text{树桩个数})$$

把这条直线画在你的图上。[要画这条直线,先用此方程式预测出 $x=1$ 和 $x=5$ 的 y 值。把得到的两个 (x, y) 点画在图上,再画一条通过这两点的直线。]

- (c) 这里的两个变量间的相关系数是 $r=0.916$ 。甲虫幼虫群数的变异,有多少百分比可以用和树桩个数的直线相关来解释?
- (d) 根据你对(a)(b)(c)部分的解答,你认为数一数树桩数,是否为预测甲虫幼虫群数的快速又可靠的方法?

15.9 葡萄酒和心脏病。表 15.1 里有 19 个国家葡萄酒消耗量和心



脏病死亡率的资料，散布图(习题 15.7)显示出还算强的相关性。根据表 15.1 的数据算出，从葡萄酒消耗量来预测心脏病死亡率的最小二乘法回归直线是：

$$y = 260.6 - 22.97x$$

用这个方程式来预测以下两个国家的心脏病死亡率，其中一个国家的成人每年平均从葡萄酒中摄取 1 升的酒精，另一国则是 8 升。利用这两个结果在你的散布图上画出最小二乘法直线。

15.10 强相关但不是线性相关。习题 14.5 列出了以下某一辆车的速率(每小时英里数)及汽油里程(MPG，每加仑英里数)的资料：

速率	20	30	40	50	60
MPG	24	28	30	28	24

用速率预测 MPG 的最小二乘法回归直线为：

$$\text{MPG} = 26.8 + (0 \times \text{速率})$$

- 画出散布图并把这条直线画上去。
- MPG 和速率之间的相关系数是 $r=0$ 。这点对于能否用回归直线来预测 MPG，提供了什么信息？

15.11 葡萄酒和心脏病 在习题 15.7 和 15.9 里，你查阅了来自表 15.1 的葡萄酒消耗量和心脏病死亡率的资料。试提出一些国和国之间的差别，是可能和饮酒习惯有交叉的。还有，关于整个国家的数据，对于个人并不能提供多少信息。所以光是靠这些数据，不能当做证据，来说明你我多喝一点葡萄酒，就可以减低心脏病的风险。

15.12 相关系数及回归。如果两个变量 x 和 y 之间的相关系数是 $r=0$ ，两个变量之间就没有直线相关。事实是当相关系数为 0 时，就是最小二乘法回归直线的斜率是 0 的时候。说明一下为什么斜率为 0 就代表 x 和 y 之间没有直线相关。画一条斜率为 0 的直线，然后解释为什么在这种情形下，要预测 y 根本不必用到 x 的值。

15.13 酸雨 酸雨的研究者连续 150 个星期在科罗拉多州的旷野地



区度量了雨的酸度。酸度是用 pH 值来测量的。pH 值低代表酸度高。酸雨研究者观察到了对应时间的直线型态。他们提出报告,最小二乘法回归直线

$$\text{pH} = 5.3 - (0.0053 \times \text{周数})$$

和数据相当吻合。

- (a) 画出这条直线的图。相关性是正还是负?用一般用语来说明这个相关性的意义。
- (b) 根据回归直线来算,研究开始时(第一周)的 pH 值及结束时(第 150 周)的 pH 值各是多少。
- (c) 回归直线的斜率是多少?清楚说明此斜率对于在这旷野地区雨中 pH 值的改变提供了什么信息。

15.14 复习一下直线 佛瑞把他存的钱藏在床垫下。开始时他有妈妈给的 500 元,之后每年再存 100 元。在 x 年之后他的存款总数 y ,可以用下列公式计算:

$$y = 500 + 100x$$

- (a) 画出这个方程式的图。(选两个 x 值,比如 0 和 10。用方程式算出对应的 y 值。在坐标图上标示出这两点,再画条通过这两点的直线。)
- (b) 20 年后,佛瑞的床垫下会有多少钱?
- (c) 如果除了开始的 500 之外,佛瑞每年存 200 元而不是 100 元,那他在 x 年后的存款总数(元)要用怎样的方程式表示?

15.15 复习直线 在刚出生之后的一段时间,一只公白鼠每周恰恰增加 40 克。(这只老鼠增重得特别有规律,不过每周 40 克仍是一个合理的增重率。)

- (a) 如果该鼠出生时重 100 克,用一个方程式表示它在 x 周之后的重量。这条直线的斜率是多少?
- (b) 画出这条直线在出生和 10 周之间的这一段。
- (c) 你会不会想要用这条直线来预测该鼠 2 岁时的体重?先预测看看,再想想结果合不合理。(1 磅等于 454 克。一只大型猫重约 10 磅。)



15.16 再谈相关系数及回归 在习题 15.3 及 15.5 里, IQ 和 GPA 的相关系数以及最小二乘法回归直线的斜率都是正的。在习题 15.7 及 15.9 里,葡萄酒消耗量和心脏病死亡率的相关系数以及最小二乘法直线的斜率都是负的。这两个数可不可能一正一负?说明你的答案。

15.17 一定要画图! 表 15.2 里有统计学家安思孔(Frank Anscombe)所准备的 4 组数据,用来说明不先画图就做计算有多冒险。四组资料的相关系数和最小平方回归直线都相等到小数点后好几位。回归方程式是:

$$y = 3 + 0.5x$$

- (a) 替每一组数据画一个散布图,并把回归直线画进每一个图里。(画回归直线可以分别把 $x = 5$ 及 $x = 10$ 代入方程式,找出对应的预测 y 值。在四个图上都画出这两点,并连接这两点画出一条直线。)
- (b) 对四组数据中的哪些值,你会愿意用回归直线来预测 $x = 10$ 时的 y 值?对每一组的答案都要给予解释。

表 15.2 探讨相关系数和回归的四组数据

数据组 A											
x	10	8	13	9	11	14	6	4	12	7	5
y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68
数据组 B											
x	10	8	13	9	11	14	6	4	12	7	5
y	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74
数据组 C											
x	10	8	13	9	11	14	6	4	12	7	5
y	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73
数据组 D											
x	8	8	8	8	8	8	8	8	8	8	19
y	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

资料来源: Frank A. Anscombe, "Graphs in statistical analysis", *The American Statistician*, 27(1993), pp. 17-21.

15.18 上课有用 一项对一所州立大学一年级生上课出勤情况及成绩的研究指出,一般来讲,上课出勤率较高的学生,成绩也较高。若



上课出勤率说明了学生成绩变异的 16%，那么上课出勤率和成绩之间的相关系数的值是多少？

15.19 日渐稀少的农场人口。在美国，农场里的人口，在上个世纪当中持续稳定的减少。以下是 1935—1980 年间农场人口(以百万人计)资料：

年	1935	1940	1945	1950	1955	1960	1965	1970	1975	1980
人口	32.1	30.5	24.4	23.0	19.1	15.6	12.4	9.7	8.9	7.2

(a) 画一个散布图。用目测法画一条用来预测某年农场人口的回归直线。

(b) 延长你的直线来预测 1990 年的农场人口。结果合不合理？为什么？

15.20 很多酒。习题 15.9 中给了从每人葡萄酒摄取酒精的升数，来预测每 100 000 人中，死于心脏病人数的最小二乘法回归直线。直线是根据 19 个富国的资料得来的。方程式是 $y = 260.6 - 22.97x$ 。一个国家的人如果喝的酒多到平均一个人摄取了 150 升酒精，那么心脏病死亡率的预测值是多少。解释一下为什么这个答案不可能是正确的。说明为什么用回归直线来做这项预测不怎么聪明。

15.21 灭火英雄灭火是不是愈灭愈糟？有人说：“火灾现场的消防员人数和该场火灾造成的损害之间有很强的正相关，所以派一大堆消防员只会造成更大的损失。”说明为什么这种推断是错误的。

15.22 你的自尊心如何？较成功的人通常对自己感觉满意，也许能提高人们对自己的信心，会有助于他们在学校和生活中的表现，所以会有一段时间，提高自尊成为许多学校的目标。加州甚至成立了一个州委员会来推动这种课程。除了“自尊使得一个人在学校有好表现”之外，你能不能为高自尊和在校表现优良的相关性找出解释？

15.23 大医院是否对你不利？一项研究显示，医院的大小(用病床数 x 来估量)和病人住院天数的中位数 y 之间正相关。这并不代表你如果选一家小点的医院就可以少住院几天？为什么？



15.24 健康和财富 一篇标题为《国家的健康和财富》的文章中说到：

健康和公民平均收入之间的正相关，是国际发展中最为人所知的相关关系之一。一般认为这个相关性反应了从收入到健康的因果关系。……然而最近出现了另一个引起人兴趣的可能：即健康和收入之间的正相关，部分可以用一个反过来的因果关系解释，即健康导致财富。

说明一个国家的较高公民收入，如何可以导致较佳的健康状况。然后解释较佳的健康状况，如何可以造成较高的收入。没有简单的方法可以决定因果关系的确切方向。

15.25 数学好不好是大学学业顺利的关键吗？以下是报纸报道大学委员会对 15 941 位高中毕业生所做的某项研究的开场白：

在高中曾修习代数和几何的少数族裔学生，能顺利完成大学学业的比例和白人学生几乎一样。一项最新研究如是报道。

大学委员会主席史都华说，高中数学和大学毕业之间的相关程度“几乎不可思议”，因此他认为，“对于能否顺利完成大学学业，数学居于关键地位”。

他表示：“这些发现，为正被认真考虑中，规定全体学生都要修习代数和几何的全国政策，提供了充分的理由。”

有哪些潜在变量可能可以解释，高中时修了好几门数学课和成功完成大学学业之间的相关性？说明一下为什么把代数和几何变成必修课，对于是否能成功完成大学学业，也许影响不大。

15.26 代糖会导致体重增加吗？用代糖来当糖的取代品的人，通常体重比用真的糖的人要重。这是不是说明使用代糖会导致体重增加？替这个相关关系找出一个较可信的解释。

15.27 看电视及学业成绩。看很多电视的儿童，和电视看得少的儿童相比，平均来说成绩较差。提出一些可以解释这种相关关系的潜在变量，因为这些变量和看很多电视以及学业成绩差同时有关系。

15.28 再谈相关系数，IQ 分数和 GPA 之间的相关系数是 $r = 0.634$ (习题 15.3)。葡萄酒消耗量和心脏病死亡率之间的相关系数是 $r =$



-0.843(习题 15.7)。这两个相关系数中,哪一个显示出较强的直线相关?说明你的答案。

15.29 魔术莫扎特 1998 年时,密歇根州的卡拉马助交响乐团用以下叙述替他们的“儿童莫扎特”营做宣传:“问题:哪些学生在语言技巧的分数高出 51 分,数学高出 39 分?答案:有音乐素养的学生。”你对于“有音乐素养”导致好成绩的这个说法,有些什么看法?

15.30 计算最小二乘法直线 你学东西的时候是不是会想知道所有的来龙去脉?以下就是从 x 预测 y 的最小二乘法回归直线的公式。先找出两个变量各自的平均 \bar{x} , \bar{y} 及标准差 s_x 和 s_y , 以及二者间的相关系数 r 。最小二乘法直线的方程式是 $y = a + bx$, 此处

$$b = r \frac{s_y}{s_x} \quad \text{截距: } a = \bar{y} - b\bar{x}$$

第 14 章的例 3 有化石骨长资料的平均数、标准差和相关系数。把这些值代入上面的公式,来验证一下在最小二乘法直线方程式:

$$\text{肱骨长度} = -3.66 + (1.197 \times \text{股骨长度})$$

以下习题需要用到有两种变量统计功能的计算机或软件,要能从数据算出最小二乘法回归直线。

15.31 教授常游泳。回到习题 15.6 的游泳资料。

- (a) 验证一下该习题中所给的最小二乘法回归直线的方程式。
- (b) 假设你只知道脉搏是 152, 现在你想要预测游了多久。找出适用于这个问题的最小二乘法回归直线方程式。你的预测是什么?
- (c) 在(a)和(b)当中的直线是不同的。详细说明为什么有两条不同的回归直线。

15.32 葡萄酒对心脏病有益吗?表 15.1 有 19 个国家葡萄酒消耗量和心脏病死亡率的资料。验证一下习题 15.9 的最小二乘法回归直线方程式。

15.33 一定要画图!可能有人会怀疑表 15.2 中的 4 组很不一样的数据,是否真的会有一样的相关系数和最小二乘法直线。请证实最小二乘法回归直线正如习题 15.7 中所列出的,是: $y = 3 + 0.5x$ (至少非常近似)。



第 16 章

消费者物价指数和政府统计

最重要的政府统计

美国政府统计机构所生产的多种数据系列中，哪一种最重要？从每月当前人口调查所得到的失业率，当然有资格当选，因为失业率上升可能会影响选举结果并改变政府政策。不过我另有提议：每月的消费者物价指数(CPI, Consumer Price Index)。

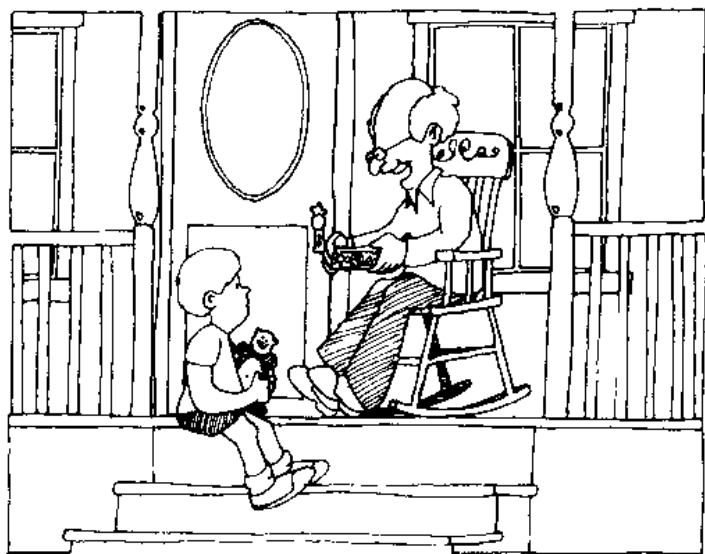
CPI 度量的是商品和服务的价格随着时间发生的变动，所以它就度量了美元随时间而下降的购买力。如果同样的商品和服务变贵了，1 美元的价值就变低了。2000 年的 1 美元能购买的东西，比 1980 年的 1 美元少，所以它其实已经是个不一样的 1 美元，虽然表面上看起



来一样。事实上 CPI 告诉我们，2000 年初的 1 美元，比起 1980 年的 1 美元来，只能买一半的东西。在 2000 年赚钱不到 1980 年两倍的人，其实购买力已经下降了。

CPI 为什么这么重要呢？它有时确实会影响选举结果和政府政策。而且，它还和经济的一大部分有直接的关系，比其他任何统计都比不上的。没有人愿意丧失购买力，所以有足够影响力的团体，会把他们的收入和 CPI 直接联系起来。于是，社会安全保险给付在 CPI 增加时会自动上升，军职和联邦文职人员的退休金亦然。而超过 200 万的工会成员有合同保障他们的工资和 CPI 连动。劳工统计局说，有超过 8000 万人的收入，直接受 CPI 影响。当 CPI 增加 1%，政府开支就自动一年增加 60 亿美元。所得税的分级标准，也会跟着 CPI 上升。你甚至可以买到价值会跟着 CPI 上升的美国储蓄公债。

CPI 对正规划未来的人也很重要。为以后的教育或退休存钱，必须考虑到美元购买力下降的问题。要比较 1980 年的美元和 2000 年的美元，我们可以利用 CPI 把它们变成有同样购买力的“真正美元”再来比较。本章将会告诉我们要如何做。



“这是真正的 1980 年的一元钞票，以后发行的就再也不一样了。”



我们大家都注意到职业运动员的薪水很高。举例来说，大联盟棒球的平均薪水，就从1980年的143 756美元，上升到了1999年的1 567 873美元。这可是跃升了一大截，不过事实上并没有表面看起来上涨这么多。1999年的1美元能够买到的东西，不如1980年的1美元多，所以1980年的薪水不能直接和1999年的比较。美元的购买力随着时间稳定下降，这个无法否认的事实等于告诉我们，每当我们比较不同年度的美元价值时，都必须事先做调整。调整很容易做。不容易的部分是怎样度量美元不断改变的购买力。政府的“消费者物价指数”就是我们需要的工具。

指数

CPI是另一种数值描述：指数(index number)。对任何在不同时间重复度量的数量变量，我们可以给此变量加上指数。指数主要是要提供变量的变化情况，它所提供的信息和下面的这句话所说的很类似：“在1990年到1996年之间，平均单日住院费用涨了46%。”也就是说，指数描述的是：从基期(base period)起算，改变量的百分比。

• 指数

指数(index number)度量的是，以变量在某个基期的值为标准，该变量值相对于基期值的比值大小。要算出变量任一值所对应的指数，可用下式：

$$\text{指数} = \frac{\text{变量值}}{\text{基期值}} \times 100$$

例1 指数的算法

1加仑的无铅汽油在1990年1月时卖1.042美元，2000年1月卖1.301美元（这些是根据美国劳工统计局搜集的资料得来的全美国平均价格）。以1990年1月为基期，2000年1月的无铅汽油价格指数为：



$$\begin{aligned}\text{指数} &= \frac{\text{价格}}{\text{基期价格}} \times 100 \\ &= \frac{1.301}{1.042} \times 100 = 125\end{aligned}$$

而基期 1999 年 1 月的无铅汽油价格指数是

$$\text{指数} = \frac{1.042}{1.042} \times 100 = 100$$

要了解指数的意义，必须知道基期是什么。因为基期的指数一定是 100，要指明 1990 年为基期，常会用“1990 = 100”的方式表示。在有关 CPI 的新闻报道中，你会看到神秘的方程式：“1982—1984 = 100”。这是一种缩写方式，代表 CPI 的基期是从 1982 到 1984 年。指数只不过是把当前值用基期值的百分比表示出来。指数 125 代表当前值为基期值的 125%，也就是比基期值增加了 25%。指数 80 代表当前的值是基期值的 80%，也就是减少了 20%。

固定市场总览物价指数

表面上看起来，指数差不多就是将简单叙述用复杂语言来伪装的一种计谋。为什么要说“消费者物价指数(1982—1984 = 100)在 2000 年 1 月时为 168.8”，而不说“消费者物价指数从 1982—1984 期间的平均到 2000 年 1 月，增加了 68.8% 呢”？事实上，指数这个字代表的意恩，通常指的不仅仅是以基期为标准所改变的量的度量。指数也告诉我们：我们在考虑的到底是什么样的变量。该变量事实上是好些数量的加权平均(weighted average)，其中权重(weight)是固定的。我们用简单的物价指数来说明这个概念。



例 2 山区居民物价指数

史密斯住在山中的小屋，力求自给自足。他只买盐、煤油，以及雇用一位焊接工。下面就是史密斯在 1990 年(基期)的全部购买情况。最后一栏中他所花的费用，是单位价乘上他所购买的单位数。

商品或服务	1990 年数量	1990 年单价	1990 年费用
盐	100 磅	0.50 美元/ 每磅	50.00 美元
煤油	50 加仑	1.00 美元/ 每加仑	50.00 美元
焊接	10 小时	14.00 美元/ 每小时	140.00 美元
			全部费用 = 240.00 美元

史密斯在 1990 年所购商品及服务的总费用是 240 美元。要算出 2000 年的“山区居民物价指数”，我们用 2000 年的价钱来计算同样的商品及服务在 2000 年的总费用。以下是计算过程：

商品或服务	1990 年数量	2000 年单价	2000 年费用
盐	100 磅	0.80 美元/ 每磅	80.00 美元
煤油	50 加仑	1.00 美元/ 每加仑	50.00 美元
焊接	10 小时	23.00 美元/ 每小时	230.00 美元
			全部费用 = 360.00 美元

同样的商品和服务，在 1990 年时花费 240 美元，2000 年时花费 360 美元。所以，2000 年的山区居民物价指数(1990 = 100)是：

$$\text{指数} = \frac{360}{240} \times 100 = 150$$

例 2 中提出的观点是，我们追踪在不同的时间，同一组商品和服务的价格。有可能史密斯在 2000 年时不愿意再雇那位焊接工，因为他工资涨得太厉害。可是没关系——计算指数用的是 1990 年的数量，完全不管史密斯在 1990—2000 年间的购买情况有没有改变。我们把当做价格追踪对象的全部商品和服务统称为市场总览(market basket)。算出



的指数就是固定市场总览物价指数(fixed market basket price index)。

• 固定市场总览物价指数

固定市场总览物价指数(fixed market basket price index)是根据一组特定的商品和服务的总价所算出来的指数。

固定市场总览物价指数背后的基本概念是：每一单项(盐、煤油、焊接工)的权重固定，不随时间改变。CPI 在本质上就是一种固定市场总览物价指数，其中包括几百种单项，代表了所有的消费行为。因为我们比较的是完全一样的项目在不同时间的价格，所以市场总览固定后，才能合理比较价格。不过我们稍后会看到，这也给 CPI 带来严重的问题。

如何使用 CPI

现在就把 CPI 想成是美国消费者买的所有东西的价格指数。2000 年 1 月的 CPI 是 168.8，代表的意思是我们在 1982 年基期花 100 美元买的货品和服务，在 2000 年 1 月必须花费我们 168.80 美元。利用对“所有东西的价格”的指数，让我们能够把不同年度的金额，转换成同一年度的美元，来进行比较。比如说，你可以在《美国统计精粹》里找到诸如标题为“住户中位收入，以 2000 年恒值美元 [Constant (2000) Dollars] 表示”之类的表。那个表已经将所有收入转换成和 2000 年美元有同样购买力的美元来表示。留意恒值美元以及真正收入这些用语。出现这种用语时，即便是在讨论不同的年度，全部的美元所代表的购买力都是一样的。

表 16.1 里有 1915—1999 年之间的每年平均 CPI。图 16.1 是用表中 CPI 值画出的线图。可以看出 20 世纪是一个通货膨胀的时代，也就是说在整个世纪当中价格一直上升，而 1973 年之后上升尤其迅速。面对这项令人沮丧的事实，在讨论美元时，若不为它们下降的购买力做些调整，就相当不明智了。以下是将某一年美元转换成另一年美元的公式。

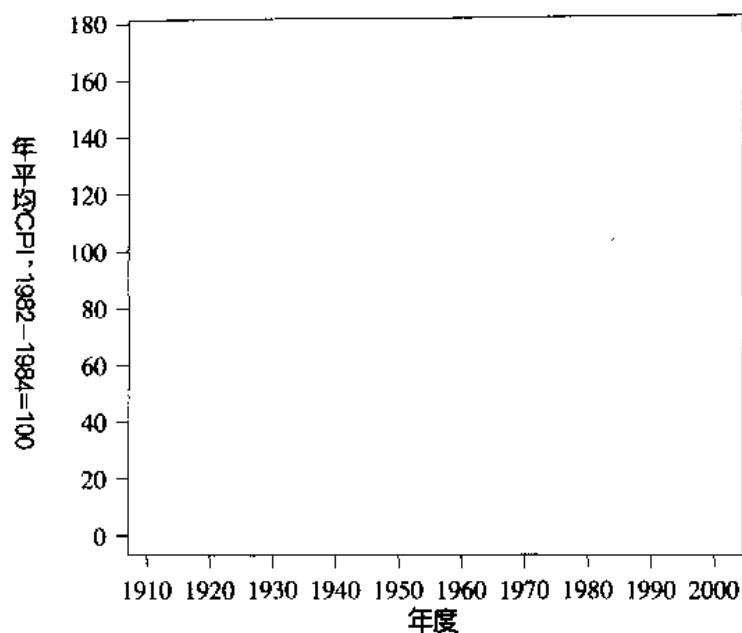


图 16.1 1915—1999 年的 CPI(1982—1984 = 100)。在 20 世纪当中, 美国的消费者物价急剧上升

表 16.1 年平均消费者指数, 1982—1984 = 100

年度	CPI	年度	CPI	年度	CPI	年度	CPI
1915	10.1	1965	31.5	1982	96.5	1991	136.2
1920	20.0	1970	38.8	1983	99.6	1992	140.3
1925	17.5	1975	53.8	1984	103.9	1993	144.5
1930	16.7	1976	56.9	1985	107.6	1994	148.2
1935	13.7	1977	60.6	1986	109.6	1995	152.4
1940	14.0	1978	65.2	1987	113.6	1996	156.9
1945	18.0	1979	72.6	1988	118.3	1997	160.5
1950	24.1	1980	82.4	1989	124.0	1998	163.0
1955	26.8	1981	90.9	1990	130.7	1999	166.6
1960	29.6						

资料来源: 美国劳工统计局。

• 为购买力的改变做调整

要将时间 A 的某美元数目, 转换成在时间 B 时有同样购买力的美元数目, 可用下列公式:

$$\text{时间 B 的美元数目} = \text{时间 A 的美元数目} \times \frac{\text{时间 B 的 CPI}}{\text{时间 A 的 CPI}}$$



请注意，你要转换到的年度的 CPI，在公式中 CPI 的比值当中，是出现在分子上。下面有些例子。

例 3 职业运动员的薪水

大联盟棒球队的平均年薪，从 1980 年的 143 756 美元，上涨到 1999 年的 1 567 873 美元。实际上到底增加了多少？我们来把 1980 年的薪水转换成 1999 年的美元。表 16.1 里有我们需要用的年平均 CPI。

$$\begin{aligned} 1999 \text{ 美元数目} &= 1980 \text{ 美元数目} \times \frac{1999\text{CPI}}{1980\text{CPI}} \\ &= 143\,756 \text{ 美元} \times \frac{166.6}{82.4} \\ &= 290\,652 \text{ 美元} \end{aligned}$$

也就是说，在 1980 年用 143 756 美元可以买到的东西，在 1999 年要花 290 652 美元。现在我们可以拿以 1999 年美元所算出来的 1980 年平均年薪（相当于 290 652 美元），来和 1999 年的实际平均年薪 1 567 873 美元做比较。即使已为现今美元较不值钱的事实做了调整，当今的运动员赚的钱还是比 1980 年的运动员高出许多。（当然平均年薪被少数几个明星球员的超级高薪给拉高了，因为 1999 年的中位薪水只有 495 000 美元。）

例 3 收入增加了？

若要讨论比较严肃的例子，我们离开职业运动员这个受娇宠的群体，来看看一般的美国人的收入情况。所有美国住户的 1980 年中位年收入是 17 710 美元。到了 1999 年，中位年收入上升到 40 816 美元。从数字看来，是 1980 年的两倍多，但是我们很清楚，上升的部分大半只是个错觉，因为美元的购买力一直在下降。要比较



这两项收入，必须把它们转换成同一年度的美元。让我们把 1980 年的中位住户收入，转成 1999 年的美元：

$$1999 \text{ 美元数目} = 17\,710 \text{ 美元} \times \frac{166.6}{82.4} = 35\,807 \text{ 美元}$$

真正的住户收入在 1980—1999 年的 19 年间，只不过从 35 807 美元上升到了 40 816 美元，上升的百分比是 14%。

但顶级薪水的人可就不同了。收入在前的 5% 住户，在 1980 至少赚了 51 500 美元。换成 1999 年美元就是：

$$1999 \text{ 美元数目} = 51\,500 \text{ 美元} \times \frac{166.6}{82.4} = 104\,125 \text{ 美元}$$

事实上，最高 5% 的住户在 1999 年的收入至少有 142 021 美元，也就是说，赚钱最多的阶层真正收入增加了 36%。

最后来看看从事生产的工人，也就是传统的男女工人。他们的平均每小时工资在 1980 年是 6.66 美元，1999 年是 13.28 美元。把 1980 年的工资重新用 1999 年的美元表示的话：

$$1999 \text{ 美元数目} = 6.66 \text{ 美元} \times \frac{166.6}{82.4} = 13.47 \text{ 美元}$$

从事生产的工人实际上的收入，在 1980—1999 年之间，反倒还下降了一些。

例 4 说明了如何利用 CPI 来比较不同年度的美元金额，展露原本隐藏的真相。以我们的例子来说，真相就是：1980—1999 年代经济繁荣的果实，大部分都被顶级收入的人享用了。换个方式来说，有特殊技能和受教育多的人得到的待遇，比从事生产的工人好得多，后者通常没有特别技能也未受大学教育。经济学家提出数种可能原因：“新经济”让知识变成利多、大量外来移民使得较不需技巧的工作有更多的人竞争、来自其他国家的竞争更多等等。至于到底为什么技能和教育的“回收”会增加这么多，以及对于受教育少的人的收入停滞状况应该做些什么，仍然是具争议性的问题。



了解CPI的意义

CPI的概念是,它是美国消费者买的所有东西的价格指数。但这项概念还需要好好调整一番才能实际应用,而且有很大部分都要用到大型抽样调查的结果

指数里面包括了谁?最常用的CPI(也有其他的,但是我们不考虑)官方名称是“城市消费者物价指数”(Consumer Price Index for All Urban Consumers)。CPI的市场总览代表居民的购买内容。“城市”的官方定义很广,所以包含了大约80%的美国人口。可是如果你住在农场里,CPI对你就不适用。

市场总览怎么选呢?不同的住户会买不同的东西,所以我们要怎么决定一个市场总览呢?由抽样调查来决定。消费者消费调查(The Consumer Expenditure Survey)搜集了29 000个住户的详细消费资料。美国劳工统计局把消费分为诸如“新鲜水果和蔬菜”、“新车及已使用的车辆”及“医院及相关服务”。然后他们会选择特定项目,比如以“新鲜橘子”,来代表市场总览中的一个类别。市场总览中的各个项目在指数计算中会有一个权重,代表该类别在总花费中所占百分比。而甚至连权重和市场总览中的项目都会定时更新,以便跟上消费者改变购买习惯的脚步。所以市场总览并不真的是固定的。

价钱是怎么决定的?再做更多的抽样调查来定。美国劳工统计局必须每个月都调查“新鲜橘子”的价钱。橘子价钱在各个城市都不同,即便在同一个城市,在不同的店价格也不一样。“购买点调查”(The Point of Purchase Survey)涵盖的16 800住户,可以让劳工统计局对于消费者到哪里去购买每一类别的商品和服务(超市、便利商店、廉价商店等等),掌握最新的资料。每个月美国劳工统计局都会在85个城市里的一些样本店里,记录下80 000项价格,而选出的样本店是可以代表消费者的实际购买习惯的。

CPI是否可以度量生活费用的改变情形?“固定市场总览物价指数”度量的是每年都过得完全一样时价格的变化,如例2所描述的。但是事实上我们并不会一直购买同样的商品和服务。我们从黑胶唱片(LP)改成买录音带及激光唱片(CD)、然后又改成买DVD音碟。我们也不会1995年或2002年买1984年出厂的车。当价格改变时,我们会改买别的——如果牛肉变得太贵,我们就少买牛肉而多买鸡肉或

你以为这就叫
通货膨胀?

当1973年因油价上涨而引发一波通货膨胀,使得之后的十年间,CPI几乎变成两倍,对此美国人很不高兴。其实这实在不算什么。阿根廷在1989年7月时,一个月之内价格上涨了127%。土耳其的里拉(lira)从1970年的14里拉兑1美元,变成2000年的579 000里拉兑1美元。1920年1月时,65个德国马克(mark)可兑换1美元,到1923年11月,则是42 000亿马克兑1美元,这才叫通货膨胀。



多买豆腐。固定市场总览物价指数没法度量生活费用的改变情形。

美国劳工统计局很努力地不断更新市场总览的内容，并针对品质的改变做调整。比如说，每一年劳工统计局都得决定，新车价格的涨幅当中，有多少是因为产品的品质提高，剩下的才在计算CPI时当做真正的涨价部分。在1967年12月—1994年12月之间，车价实际上涨了313.4%，但是CPI中的新车价格只上升了172.1%。在1995年，经过对较佳品质做调整之后，使得商品和服务价格的整体上涨，从4.7%降到了只有2.2%。商品和服务的价格，构成CPI的70%，其余的大部分是居住费用：租公寓或者买房子。房价是美国劳工统计局的另一个问题。人们买房子一方面是要住，另一方面也认为拥有房子是一项好的投资。如果我们肯多花钱买一栋房子是因为认为它是一项好投资，就不应把整个房价算进CPI。

说到现在已经很清楚了，CPI并不是固定市场总览物价指数，虽然这是考虑它的最好出发点。当新产品出现和我们的购买习惯改变时，劳工统计局必须不断地更新市场总览。他们还得调整抽样调查所得到的价格，来把较好的品质以及房价的投资部分等因素考虑进去。然而CPI还是不能度量我们生活费用的变化，比如说它就没有把税金考虑进去，而税金当然是我们生活费用的一部分。

即使我们取得共识，让CPI只考虑我们购买的商品和服务，它还是不能量出我们生活费用的变化。照道理讲，真正的“生活费用指数”(cost of living index)应该度量同样生活水平的花费随时间而变化的情况。这就是为什么我们先从固定市场总览物价指数谈起，因为它也度量了过同样生活的费用所改变的情形，只是把情况简化，将“同样”解释成购买完全一样的东西。如果我们为了省钱改买豆腐不买牛肉，而仍然心满意足，我们的生活水平就没有改变，此时生活费用指数应该不计入牛肉和豆腐的价差。如果我们愿意多花钱去买环保产品，则我们是为了较高的生活水平而付费，所以指数处理这项花费，就应该像对待较佳品质的新车一样。美国劳工统计局说，他们希望能够让CPI来追踪生活费用改变的情形，但是在真实世界中，不可能出现真正的生活费用指数。



统计学上的争议

CPI 把通货膨胀夸大了吗?

1995 年美国联邦储备委员会 (Federal Reserve) 主席格林斯潘 (Alan Greenspan) 估计, CPI 每年把通货膨胀约高估了 0.5% 到 1.5% 左右。格林斯潘先生对此颇为不悦, 因为 CPI 上升会自动增加联邦支出。1996 年底, 一群由参议院财政委员会指派的外界专家估计, CPI 在过去差不多每年都把通货膨胀高估了 1.1%。美国劳工统计局承认, CPI 的确高估了通货膨胀, 但是认为专家说的每年 1.1% 太多了。

CPI 显示出的美元价值下跌速度比实际上的要快, 其原因一部分是由于 CPI 的本质, 一部分是由于美国劳工统计局对 CPI 背后的巨大机制的细节做调整时, 有其速度上的限制。先来想想这些细节, 诸如数码相机及电脑平面屏幕这类新产品的价格, 通常刚上市时很贵, 但之后会迅速下跌。CPI 的市场总览更新得太慢, 体现不出这些较早的降价。而且 CPI 把价格较低的廉价商店纳入样本的速度也太慢。虽然劳工统计局很努力地在为较佳品质做调整, 外界专家仍然认为这些调整通常显得不足而且又太慢。但美国劳工统计局在这些细节上仍做了许多改进, 改良后的 CPI 原可使 1978—1988 年之间的实

际 CPI 每年少成长 0.5%。

更大的议题是 CPI 本质上就是固定市场总览物价指数。这类指数本就容易偏高, 因为当消费者试着通过购买当月较便宜的产品来维持生活品质, 因此从买牛肉改成买豆腐, 或者又改回来时, 指数是体现不出这些转变的。

外界专家对于 CPI 的批评, 说得明白一点也就是 CPI 不能追踪“生活费用”这件事, 他们所做的第一项建议是: “美国劳工统计局应该设法建立一项生活费用指数, 当做度量消费者物价的目标。”美国劳工统计局对此原则上表示同意, 但是不论是他们自己, 或者任何其他人, 都不知道实际上要如何做。该局还说: “对于‘生活质量’改变的量度, 可能要依靠过多的主观判断, 以致无法提供能令人接受的 CPI 调整依据。”即便如此, 有一种在原则上和生活费用改变的度量比较接近的指数, 将于 2002 年出现。请拭目以待 CPI 的这些改变, 以及随之而来的抱怨, 因为指数增加得较慢, 会使诸如领社会安全保险金的人收入减少。



政府统计的处境

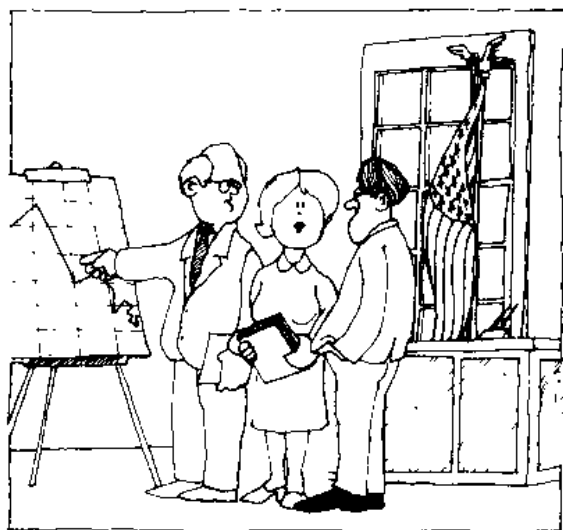
现代化国家的运作依赖统计。经济方面的资料尤其可以引导政府政策,并为私人企业及个人决策提供资料。物价指数、失业率及其他许多较不常见的系列资料,则是由政府的统计机构整理出来的。

有些国家有单一的统计机构,比如加拿大统计局(Statistics Canada, 网址 www.statcan.ca)。其他有些国家的统计单位则较小,附属于政府的各部门。美国是个特例:共有 72 个联邦统计单位,但彼此之间却很少联系。普查局和劳工统计局是最重要的两个,不过你偶尔可能会用到经济分析局(Bureau of Economic Analysis)、国家卫生统计中心(National Center of Health Statistics)、司法统计局(Bureau of Justice Statistics)或其他属于联邦政府的统计单位的资料。

某些国家统计机构的领导对政府统计机构做的 1993 年排名,把加拿大摆在首位,而美国和英国、德国并列第 6 位。头几名的国家通常都拥有单一且独立的统计机构。1996 年时,英国将它的一些主要统计机构做了调整,成为新的国家统计局(Office for National Statistics, 网址 www.statistics.gov.uk)。而美国政府的统计单位仍呈分裂状态。

老百姓需要政府统计单位提供什么呢?首先,他们需要正确、及时并跟得上社会及经济的变化脚步的资料。要快速整合出正确的资料需要相当大量的资源。想一想产生失业率及 CPI 的大型抽样调查就知道了。美国的几个主要统计单位以正确性出名,他们将资料公诸于社会大众的迅速程度也领先全世界,但他们在“跟上变化”这方面的记录就不那么好了。为了购物习惯和品质的改变,他们必须努力调整 CPI,这是一个问题;而另外一个问题是:美国经济统计根本跟不上许多趋势,诸如经济活动的重心已从制造业转向服务业。对于我们经济资料的整体状况,企业界已表示过强烈的不满。

很多的困难源于缺钱。在 1980 年之后,减少联邦支出已是美国政治上的优先项目。政府统计机构人员减少,计划也缩减,而薪水低就难以吸引最好的经济学家和统计学家为政府工作。政府应该花多少钱在数据上面,也和我们对于政府应该生产哪些数据的看法有关。确切一点来说,政府应该产生主要为私人企业所用,而不是政府自己的决策者所需的资料。也许这些资料该由私人公司处理,或者只提供给愿意付费的人。这已经是政治哲学而不是统计问题,但是有助于决



“好，我知道我们必须了解经济的走向，不过我们有必要公布这些统计结果，搞得尽人皆知吗？”

定，我们愿意政府花钱来做什么等级的统计。

对于政府统计来说，不受政治影响跟正确性与及时性一样重要。当统计单位隶属政府的某个部时，就可能受该部的需要及期望影响。美国普查局隶属商业部，是为企业服务的。美国劳工统计局隶属劳工部，因此企业和劳工都有“自己的”统计单位。这些统计单位的专业人士很成功的抗拒了直接的政治干扰——比如说，不好的失业率报告从来没有刻意留到选举之后才发表。但是间接的干预明显存在，比如说，美国劳工统计局必须和美国劳工部的其他活动争预算。当美国国会拒绝让普查局用抽样调查来补救 2000 年普查的不足时，可以看出政治干预统计工作的情况似乎是在增加中。

1996 年英国重组统计单位，促成这项举措的部分原因是，大家普遍认为政治影响力太强了。比如说，在 20 世纪 80 年代，英国把失业率如何度量的细节改了好多遍，而且几乎每一次的改变都有把公布的失业率降低的效果，而这正是政府乐于见到的。

我赞成单一的“美国统计局”(Statistics USA)不隶属于任何其他政府部门，就如加拿大一样，如此的整合因为减少了重复性，也可能有助于减少经费，至少可以对于哪些计划应该分到有限资源当中的较大部分，做统一的规划。虽然整合似乎不太可能，但是加强各个联邦统计机构之间的协调，也可以大致达到相同的目标。



社会统计的问题

在政府、媒体或一般大众心中，已确认了全美经济统计的地位。美国政府也整合了许多社会议题方面的资料，诸如教育、健康、住宅及犯罪等方面，然而社会统计仍不如经济统计完整。对于人们花多少钱买食物，我们有很好的资料；但是对于有多少人营养不良，资料则少得多。社会资料的整合方式也不如经济资料的严谨小心。一般来说，经济资料是根据较大的样本取得，资料搜集得较频繁，公布的时间间隔也较短。理由很简单：每个月政府都要用经济资料来引导经济政策。社会资料帮助我们了解社会以及社会问题，但是并没有任何短期处理的必要。

还有其他原因使政府不太想搜集社会资料。很多人都不喜欢政府询问个人的性行为或宗教的问题。很多人认为，政府不应该问我们怎么想，问“你上次什么时候看医生？”是可以的，但是不可以问“你对于医疗质量的满意程度如何？”不愿回答这类问题的态度，反映出一般人怀疑政府介入过多。然而像性行为模式和艾滋病的散播有关，这和对医疗质量是否满意等议题，对老百姓都是很重要的。关于这些议题的事实和意见，都可能影响选举及政策的制定。我们要怎么样可以不断搜集资料，取得对一些社会议题的准确信息，而同时又可以让政府陷入生活、宗教信仰及其他一些敏感主题呢？

美国的解决方法是由政府补助大学做社会调查。比如，政府曾经一度决定要举行抽样调查，询问有关人们性行为的问题，以做为制定艾滋病政策的部分资料，后来却打了退堂鼓。结果政府资助了芝加哥大学的美国全国民意调查中心做规模小得多的民意调查，对象只有3 452位成人。NORC的全面社会调查由政府的国家科学基金会(National Science Foundation)资助，和当前人口调查及形成CPI基础的那些样本一样，都列名于任何美国最重要抽样调查的名单上。GSS包含“事实”和“意见”两类项目：受访者回答的问题包括工作稳定性、工作满意度，以及对所住城市、朋友及家庭的满意程度，另外也谈关于种族、宗教和性的问题。如果政府问到去年是否看过限制级电影，很多美国人会反对，但是当GSS问这个问题时，他们就会回答。

这种政府出资，而由大学主导的抽样调查的间接系统，照顾到了



美国人觉得政府不应该过度侵犯个人的感觉。这种间接系统也让调查绝缘于大部分的政治压力。但是政府的裁减预算也延伸到了 GSS, 现在 GSS 这个自称是“几乎每年做”的调查, 也由于缺钱已经好几年没有抽样了。其实我认为 GSS 是很划算的。

网络寻奇

消费者物价指数在网络上栖身于美国劳工统计局的网站。

<http://www.bls.gov/cpi/>。查查“常见问题”(Frequently Asked Questions), 可以看到美国劳工统计局对 CPI 的解释及使用情形。最新的 CPI 出现在贴在此网站上的最近一次发布的新闻稿。

如果你喜欢数据, 就去美国劳工统计局(BLS)的数据页:
<http://www.bls.gov/data/>, 在 Price & Living Conditions 下点击 CPI Average Price Data, 自己瞧瞧像白吐司和汽油这类东西的价格有些什么样的变化。

美国政府的统计机构分散在各处, 但他们曾合力建了一个 Fed-Stats 网站, 通过此网站可链接上所有这些机构。只欣赏各式各样的美国政府统计资料的话, 就去 www.fedstats.gov, 点击 Agencies。这里蕴藏着有关美国的各式各样丰富的数据。



本章重点摘要

指数描述一个变量以某一**基期**的值为标准的对应值。**固定市场总览物价指数**是描述一整组商品和服务的总价的指数。你可以把政府的**消费者物价指数**，想成是包含消费者购买的所有商品和服务的固定市场总览物价指数。因为CPI是消费者物价如何随时间改变的一项指标，我们可以利用它把某一个年度的美元，换算成在另一年度有同样购买力的美元金额。要**真正比较**两个不同年度的美元金额，这个转换是必须的。

CPI背后的细节非常复杂。它用到好几个大型抽样调查所得到的数据。它并不是真正的固定市场总览物价指数，因为它要针对已改变的购买习惯、新产品以及改良的品质做调整。

政府统计机构生产供政府做决策以及企业和个人做决定时所需要依据的数据。数据必须准确、及时并且不受政府干预。因此，政府统计机构的能力和独立性，攸关老百姓的利益。



第16章 习题

如果你需要用到某年的 CPI, 但表 16.1 里面却没有, 就用表里面在你要的那一年之后, 最接近那一年的 CPI。

16.1 汽油价格。年底的无铅汽油平均价格变化如下:

1985 年	每加仑 1.208 美元
1990 年	每加仑 1.354 美元
1995 年	每加仑 1.101 美元

算出 1985、1990 及 1995 年的油价指数(1990 = 100)。

16.2 读大学的费用。CPI 计算大学学费的部分, 在 2000 年 1 月是 326.0(1982—1984 = 100)。那个月的整体 CPI 是 168.8。

- (a) 明确说明指数 326.0 对于在基期和 2000 年初之间大学学费的上涨情形提供了什么信息?
- (b) 大学学费涨得比起整体消费者物价还快得多。你是如何知道这点的?

16.3 汽油价格。用习题 16.1 的结果来回答下列问题。

- (a) 1985—1995 年之间, 油价指数改变了多少点? 改变了多少百分比?
- (b) 1990—1995 年之间, 油价指数改变了多少点? 改变了多少百分比?

你会发现指数改变的点数和改变的百分比, 只有当我们从基期起算时会相等, 否则不会相等。

16.4 排放有毒物质。美国环境保护局(Environmental Protection Agency)要求工业界, 只要任何有毒化学物有外泄的情形都要报告。外泄总量(单位: 千磅)在 1988 年是 3 395 867, 1995 年是 1 964 926, 1997 年是 1 941 870。以 1988 年为基期, 算出这几个年度有毒化学物的外泄指数。1988—1997 年之间, 外泄增加或减少的百分比是多少?



16.5 洛杉矶和纽约。美国劳工统计局除了公布全国 CPI 之外，也公布几个主要大都会区的个别 CPI。2000 年 1 月洛杉矶的 CPI 是 167.9，纽约是 179.2(1982—1984 = 100)。

(a) 这些数字告诉我们，在基期和 2000 年初之间，纽约的物价增加得比洛杉矶快。说明为什么我们知道是这样。

(b) 这些数字并没有告诉我们，2000 年 1 月的纽约物价比洛杉矶物价高。说明为什么如此。

16.6 食疗信徒的物价指数。一位食疗信徒只吃牛排、饭和冰淇淋。在 1990 年他去购买了：

项目	1990 年数量	1990 年单价
牛排	200 磅	5.45 美元/每磅
米	300 磅	0.49 美元/每磅
冰淇淋	50 加仑	5.08 美元/每加仑

在看望过他母亲之后，这位食疗信徒把橘子加进他的日常食物。橘子在 1990 年 1 磅卖 0.56 美元，以下是该食疗信徒在 2000 年购买的食物。

项目	2000 年数量	2000 年单价
牛排	175 磅	6.59 美元/每磅
米	325 磅	0.53 美元/每磅
冰淇淋	50 加仑	6.64 美元/每加仑
橘子	100 磅	0.61 美元/每磅

请算出 2000 年的固定市场篮子食疗信徒物价指数(1990 = 100)

16.7 古鲁物价指数。古鲁(印度教的宗教领袖)只购买橄榄油，腰布和《阿闍婆吠陀》*，他从阿闍婆吠陀中选出曼特罗诗歌送给他的信徒。以下是他在 1985 和 1995 年购物的数量和价格。

* 译注：一种经文。

项目	1985 数量	1985 价格	1995 数量	1995 价格
橄榄油	20 品脱	2.50 美元/每品脱	18 品脱	3.80 美元/每品脱
腰布	2	每件 2.75 美元	3	每件 2.80 美元
《阿闍婆吠陀》	1	10.95 美元	1	12.95 美元



根据这些数据,请找出1995年的固定市场总览占鲁物价指数(1985=100)。

16.8 班比诺诅咒(Curse of the Bambino) 1920年波士顿红袜队以125 000美元把贝比鲁斯(Babe Ruth)转让给纽约洋基队。之后在1920年和2000年之间,洋基队赢了26次世界杯而红袜队一次也没赢。1920年的125 000美元在1999年相当于多少美元?

16.9 梦想 茱莉在1995年开始读大学时定了个目标,希望一毕业就有35 000美元的年薪。她于1999年毕业时必须赚多少钱,才能拥有和1995年的35 000美元一样的购买力?

16.10 活太久了?若夫妇两人65岁时都健在,则其中有半数的夫妇,他们之中至少有一人会活过93岁,即再活28年之久。你听了应该会感到吃惊。莫娜和比尔于1970年退休之后,收入为每年10 000美元。他们过得不错——这个收入差不多是1970年时的中位家庭收入。28年之后,也就是1998年时,他们必须有多少收入才能维持同样的购买力?

16.11 微波炉大甩卖 新产品的价钱通常一开始很高,然后迅速下降。最早的家庭用微波炉于1955年时售价1 300美元,而你现在可以用100美元买到更好的微波炉。先去找出最新的CPI(美国劳工统计局的网站上有),然后利用这个CPI把现在的100美元转换成1955年的美元。把得到的数字和1 300美元相比,看看微波炉的价格真正降了多少。

16.12 高尔夫球高手。1999年老虎伍兹在职业高尔夫联盟巡回赛共赢了6 620 970美元的奖金。而1938年的奖金王是斯尼德(Sam Snead),他共赢223 274美元。在1980年领先的华生(Tom Watson),在那年共赢了1 041 002美元。这几项金额真正价值的高下如何?

16.13 迪马杰奥。洋基队中外野手迪马杰奥(Joe DiMaggio)1940年的薪水为32 000美元,1950年为100 000美元。把他在1940年的薪水用1950年的美元表示。



16.14 打电话到伦敦。经由 AT & T 公司打 10 分钟长途电话到伦敦，在 1976 年时要花 12 美元，而在 1999 年是 11 美元，请问从 1976 年到 1999 年之间真正的价格降了几个百分点？

16.15 读哈佛的花费 哈佛大学在 1976 年对于学费和食、宿共收取 5 900 美元，到了 1999 年则变成 32 164 美元。把哈佛在 1976 年的收费用 1999 年的美元表示。读哈佛的花费上涨得比整体 CPI 快还是慢？你怎样知道的？

16.16 最低工资 美国联邦政府对于雇主付给工人的待遇，有最低工资(时薪)的规定。工人希望最低工资订得高些，但许多经济学家认为，最低工资过高，会使雇主不愿雇用技能差的工人，而导致失业率增高。

以下是美国联邦最低工资的变化情形，单位是每小时美元。

年度	1960	1965	1970	1975	1980	1985	1990	1995	1999
最低工资	1.00	1.25	1.60	2.10	3.10	3.35	3.80	4.25	5.15

用表 16.1 的年平均 CPI，把最低工资重新用 1960 年恒值美元表示。在同一组坐标轴上画两条线图，一条画出这些年实际上的最低工资，另一条画出用恒值美元表示的最低工资。假设你要解说的对象是懂统计的人，详细说明你的图显示出怎样的最低工资变迁状况。

16.17 大学学费。普度大学对印第安纳州居民收取的学费增长情形如下(如果你有资料，用你自己大学的学费来算，单位是美元)：

年度	1981	1983	1985	1987	1989
学费	1 158	1 432	1 629	1 816	2 032
年度	1991	1993	1995	1997	1999
学费	2 324	2 696	3 056	3 336	3 624

利用表 16.1 的年平均 CPI，把每年的学费重新用 1981 年恒值美元表示。在同一组坐标轴上画两个线图，一条代表这些年的实际学费，另一条代表用恒值美元表示的各年学费。把解说对象当作不懂统计者，说明一下你的图里有何信息。



16.18 2000年与1980年的比较。本章刚开始时曾提到“2000年年初的1美元，比起1980年的1美元来，只能买一半的东西”。2000年1月的CPI是168.8。表16.1告诉我们，1980年的平均CPI是82.4，详细说明为什么这些数字可以用来佐证上面说的“只能买一半的东西”这句话。

16.19 收入增加了？在例4中我们见到住户的真正中位收入(用1999年美元表示)，从1980年的35 807美元上升到1999年的40 816美元。跻身最高收入的5%住户的门槛(真正收入)，在该段期间从104 125美元上升到142 021美元。请验证我们在例4中所说的：中位收入上升了14%，而最会赚钱的阶层的真正收入上升了36%。

16.20 有线电视。假设国内的有线电视系统会增加频道并调升收费。即使消费者付的费用变多了，针对有线电视收费所算的CPI也不一定会上升，说明一下为什么。

16.21 CPI中的项目比重。买房子花费(去除投资部分之后)占CPI的20%。房租购成CPI的6%。这个20%和6%是哪里来的？为什么买房子占的比重较重？

16.22 CPI对我不适用。CPI不一定得出你个人所经历的价格改变状况。解释一下为什么CPI不适用于以下这些人？

- (a) 住在蒙大拿州养牛的牧场上的玛夏。
- (b) 用老式炉子取暖，也没有冷气的古姆。
- (c) 去年出了严重车祸，所以大半时间在康复中心度过的路易和玛莉亚。

16.23 季节调整。就像许多政府数据系列一样，在CPI公布时，有未经调整和经季节调整两种形式。美国劳工统计局声明，他们“强力推荐以没有因季节变动做调整的指数(即未经季节调整之指数)为调升的指标。”此处的“调升”指调整薪水或其他给付以跟上CPI的变化。为什么这里用未经调整的CPI比较合适？

16.24 CPI续论。除了全国CPI之外，美国劳工统计局还公布4个区域(region)和26个局部地区(local area)的CPI。每个区域或地区



CPI, 是根据全美国价格样本中, 发生在该区域或地区的部分而得到的。美国劳工统计局说使用这些局部 CPI 时要很小心, 因为它们比起全美国或者区域的 CPI 来, 变化要大得多。为什么会这样?

16.25 贫困户标准。美国联邦政府每年都公布不同人口数的住户之“贫困户标准”。收入低于该标准的住户, 就被视为贫困户。某位经济学家查阅了历年的贫困户标准, 并宣称这些标准“显示出一种形态, 也就是在一般大众的真正收入上升时, 真正的贫困户标准也在上升。”“真正的贫困户标准上升”这个说法, 等于在说官方发表的贫困户标准怎样?

16.26 真正的薪水 在多篇有关美国停滞不动的薪水的报道中的一篇提到: “在 20 世纪 80 年代, 几乎所有收入群都面对了真正薪水缩水的局面。然而收入在第 33 百分位数的在职者, 真正薪水降了 14%, 第 66 百分位数的在职者, 只降了 6%, 而在分布顶端的在职者, 反倒还升了 1%。”

(a) “收入分布的第 33 百分位数”是什么意思?

(b) “真正薪水”是什么意思?

16.27 可以省钱? 要在政府统计上面省钱的一个方法, 是把样本都变得小一点。比如说, 我们也许可以把当前人口调查从 50 000 住户减为 20 000 住户。向不懂统计的人详细说明, 为什么这样缩小样本会使所得数据的精确性减低。

16.28 全面社会调查。全面社会调查的一大重点, 是年复一年问许多相同的问题。你认为这样做的目的是什么?

16.29 度量犯罪案件造成的影响。我们希望能在社会统计中, 加入犯罪案件多寡的量度以及罪行对人们态度及行为造成的影响的量度。对以下各项各提出一些可能的量度。

(a) 从诸如警方纪录等官方来源搜集的统计资料。

(b) 对老百姓抽样调查所得到关于事实的信息。

(c) 用抽样调查方式得到的关于意见和态度的信息。

16.30 统计机构。替底下每个美国政府统计机构的工作做一个简短



描述。要找相关信息可以到 FedStats 网站 (www.fedstats.gov)，再点击 “Agencies”。

- (a) 美国商业部经济分析局 (Bureau of Economic Analysis, Department of Commerce)。
- (b) 美国国家教育统计中心 (National Center for Education Statistics)。
- (c) 美国国家卫生统计中心 (National Center for Health Statistics)。

第二部分 复习

数据分析是一门用图和数值摘要来描绘数据的艺术。数据分析的目的，是要让我们能看到并了解一组数据的重要特色。第 10 章谈了基本的图，特别是饼状图、柱状图和线图。第 11 章到 13 章提出了描述一个变量的分布要用到的统计观念和工具，以说明数据分析背后的概念。

图 II.1 把这些重要概念给组织起来。我们先用数据画图，然后用平均数和标准差或者五数综合来描述分布的中心和离度。最后一步，就是用正态曲线当做整体形态的模型，而对数据做了很具体的综合，不过这一步只对某些数据适用。最后两步中的问号是要提醒我们，数值摘要和正态分布的用处有多大，是要看我们在检视数据的图时有什么发现而定。不规则的形状，或者明显分成几个群的数据，是不能只用简短的综合数值或模型来描述的。

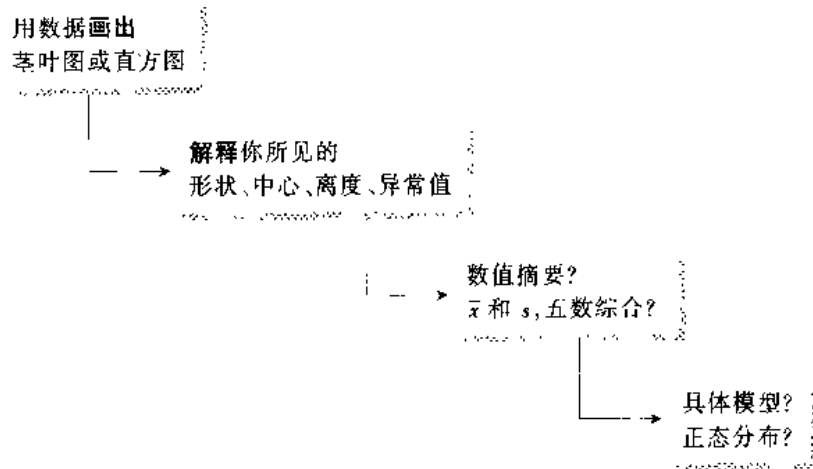


图 II.1

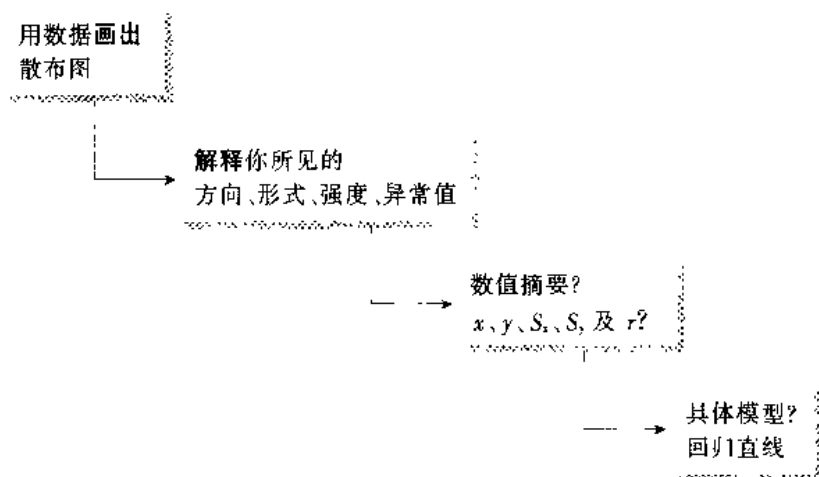


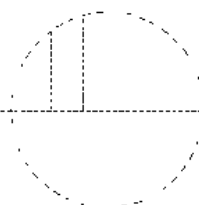
图 II.2

第 14 和 15 章把同样的概念用在两个数量变量的相关关系上面。图 II.2 把图 II.1 的重要观念重新呈现一遍，只是细节改成符合现在讨论的情境。我们一定是从用数据画图开始。如果画的是散布图，我们学到的数值摘要只适用在散布图上大致呈现直线形态的数据。此时的数值摘要，就是两个变量分别的平均数、标准差以及二者之间的相关系数。在散布图上画出的回归直线，对整体形态提供了一个具体模型，我们可以用它来预测。最后两步当中又画了问号，提醒我们相关系数和回归直线只能用来描述直线相关。

相关关系常常引出因果问题。我们知道，得自随机化比较实验的证据，是要决定一个变量导致另一个变量改变的“黄金标准”。第 15 章有更多的讨论，提醒我们即使没有直接因果关系，数据还是可



能显示很强的相关性。我们永远要考虑到，隐藏在背后的变量可能会有什么样的影响。在 16 章中我们见到一种新的描述：“指数”，其中“领衔”的例子是消费者物价指数。16 章也谈到政府统计机构，这是统计世界中较安静但重要的一部分。





第二部分 重点摘要

以下是你读了第 10—16 章后，应该具备的重要技能。

A. 展示分布

1. 分辨得出类别变量和数量变量。
2. 知道何时可以用饼状图，何时不可以。
3. 会用柱状图画出类别变量的分布，或者概括来说，会比较相关的量。
4. 会解读饼状图和柱状图。
5. 会画线图，并描述数量变量随时间变化的情形。
6. 认得出线图中出现的形态，诸如趋势和季节变动。
7. 对不当的图要有警觉，尤其是象形图，还有线图中的刻度选择问题。
8. 会画单一数量变量分布的直方图。
9. 对于比较小组的观测值，会画出其分布的茎叶图。必要时会将数据四舍五入，以便画出较有效的茎叶图。

B. 描述分布(数量变量)

1. 会从直方图或茎叶图中找出整体形态，以及明显偏离整体形态的部分。
2. 根据直方图或茎叶图来评估，分布的形状是为大致对称、明显有偏还是二者皆非。判断该分布是单峰还是多峰。
3. 除了对于形状的形容之外，还能用中心和离度的数值量度来描述整体形态。
4. 判断哪一组的中心和离度的量度比较恰当：平均数和标准差（对于对称分布最适用）或五数综合（对偏斜分布最适用）。
5. 认得出异常值，并提出合理的解释。



C. 分布的数值摘要

1. 会找出一组观测值的中位数 M 及四分位数 Q_1 及 Q_3 。
2. 会写出五数综合及画箱形图：能从箱形图中找到中心、离度和是否对称等性质。
3. 能算出一小组观测值的平均数 \bar{x} 及标准差 s (用计算机)。
4. 了解中位数比起平均数，较不受极端值的影响。看得出偏斜分布的平均数会被拖离中位数，朝着长尾的方向偏离。
5. 知道标准差的基本性质： s 必定 ≥ 0 ；只有在所有观测值都相等时才会有 $s=0$ ，而观测值散得愈开， s 就愈大； s 的单位和数据原来的单位是一样的，异常值或偏斜分布会强力拉抬 s 的值。

D. 正态分布

1. 会说明密度曲线是用来描述数量变量分布的。
2. 认得出正态曲线的形状，并且可用目测法从正态曲线估计出平均数和标准差是多少。
3. 会利用 68-95-99.7 规则以及对称性，说出正态曲线下有多少百分比的观测值落在特定两点之间，在这两点都落在平均数上或者分别是在平均数的两边，距平均数 1、2 或 3 个标准差时，都能答得出来。
4. 能找出并解释一个观测值的标准计分。
5. (此题可略过) 会用表 B 找出正态分布中某一个值相当于多少百分位数以及对应某个百分位数的值。

E. 散布图与相关系数

1. 对于以同样对象度量得到的两个数量变量，会画散布图来呈现两者之间的相关关系。解释变量(如果有的话)要放在图的横轴。
2. 会描述散布图整体形态的形式、方向及强度。更具体点说，要认得出正相关、负相关及直线形态。能分辨散布图中的异



常值。

3. 会判断用相关系数来描述两个数量变量间的关系是否恰当。
会用计算机找出相关系数 r 。
4. 了解相关系数的基本性质： r 只度量直线相关的强度及方向；
 r 的值永远都在 -1 — 1 之间；只有在 100% 的直线相关时，
才会有 $r = \pm 1$ 的情况出现；当直线相关愈来愈强， r 就会离
0 愈来愈远而趋近 ± 1 。

F. 回归直线

1. 会解释直线方程式 $y = a + bx$ 中的斜率 b 和截距 a 代表什么意思。
2. 给了直线方程式，就能画出直线的图。
3. 给了回归直线，不论是图还是方程式，都能用来预测对应 x 的 y 值。了解预测若超出数据的范围时的风险。
4. 能用相关系数的平方 r^2 ，来描述在一个变量的变异中，有多少百分比可以用和另一个变量的直线相关来解释。

G. 统计及因果

1. 能对观察到的两变量之间的相关性提出合理解释，到底是：
直接因果关系，潜在变量的影响，还是二者皆有。
2. 对于因果关系的声明，能评估其统计证据是否够强，尤其是在无法做实验时也有此能力。

H. 消费者物价指数(CPI)及相关议题

1. 会计算和解释指数。
2. 对于小的市场总览，会计算固定市场总览物价指数。
3. 会用 CPI 来比较不同年度美元的购买力。会解释“真正收入”是什么意思。



第二部分 复习习题

复习习题都是很短且直截了当的题目，能帮你加强在本书每一部分中学到的基本观念和技巧。

II.1 美国各州贫困户状况。表 II.1 中有美国密西西比河以东的 26 州中，每一州符合贫困户标准的居民所占百分比。替这组数据画个茎叶图。该分布是大致对称、右偏抑或左偏？哪些州是异常值（若有的话）？

表 II.1 1997 年符合贫困户标准的各州居民百分比

州	百分比(%)	州	百分比(%)	州	百分比(%)
亚拉巴马	15.7	马里兰州	8.4	宾州	11.2
康涅狄格	8.6	马萨诸塞	12.2	罗德岛	12.7
特拉华	9.6	密歇根	10.3	南卡罗来纳	13.1
佛罗里达	14.3	密西西比	16.7	田纳西	14.3
佐治亚	14.5	新罕布什尔	9.1	佛蒙特	9.3
伊利诺伊	11.2	新泽西	9.3	弗吉尼亚	12.7
印第安纳	8.8	纽约	16.5	西弗吉尼亚	16.4
肯塔基	15.9	北卡罗来纳	11.4	威斯康星	8.2
缅因	10.1	俄亥俄	11.0		

资料来源：《美国统计精粹》。

II.2 四分卫。表 II.2 中是美国国家足球联盟(NFL, National Football League)的先发四分卫在 1999 年球季的传球总码数。(这些是在季末时的先发四分卫，而其中有些下场数较少。)替这些数据画个直方图。分布有没有明显的形状？是大致对称、明显左偏、明显右偏还是以上皆非？哪些四分卫（如果有的话）是异常值？

II.3 美国各州贫困户状况。找出表 II.1 中贫困户数据的五数综合。

II.4 四分卫。找出表 II.2 中 NFL 四分卫传球数据(码数)的五数综合。

II.5 美国各州贫困户状况。从表 II.1 的数据中，算出美国各州符合贫困户标准的居民的平均百分比。如果我们去掉密西西比州，平均百分



表 II.2 1999 年 NFL 四分卫的传球码数

四分卫	码数	四分卫	码数	四分卫	码数
艾克门	2 964	佛鲁提	3 171	卢卡斯	1 678
班克斯	2 136	佛瑞罗	2 117	曼宁	4 135
波耶览	4 436	甘能	3 840	马利诺	2 448
布雷克	2 670	贾西亚	2 544	麦克弄	1 465
布烈叟	3 985	乔治	2 816	麦克奈尔	2 179
布鲁奈	3 060	葛巴	3 389	派得森	1 276
钱德勒	2 339	葛里斯	3 032	普勒玛	2 111
柯林斯	2 316	哈堡	2 761	陶立佛	1 916
考其	2 447	詹森	4 005	汤柴克	1 625
狄尔法	1 619	奇纳	3 346	华纳	4 353
法瑞	4 091				

比会增加还是减少?为什么?算出其他 25 州的平均来证实你的答案。

II.6 大头?军方报道说,男性士兵的头围分布大致是正态分布,平均数 22.8 英寸,标准差 1.1 英寸。用 68-95-99.7 规则来回答以下问题。

(a) 最中间的 95% 头围,会落在哪两个数字中间?

(b) 头围超过 23.9 英寸的士兵占多少百分比?

II.7 SAT 分数。SAT 考试的分数经过设定,使得分数的分布大致是平均数 500、标准差 100 的正态分布。不要参考任何表格,回答以下问题。

(a) 中位 SAT 分数是多少?

(b) 你为分数在 400—600 分之间,但还想考得更好的学生设了家教班。SAT 分数在 400—600 之间的占多少百分比?

II.8 解释相关系数。你手上有一门基础统计课程学生的目前学业平均(GPA)以及他们在这科第一次考试的分数。GPA 和第一次考试分数之间的相关系数 r 度量的是什麼,请尽量具体地说明。

II.9 小鼠的数据。为了一份生物作业,你量了同种的 12 只小鼠的尾长(厘米)及体重(克)。以下每一项的量度单位各是什么?

(a) 平均尾长。



- (b) 尾长的第一四分位数。
- (c) 尾长的标准差。
- (d) 尾长和体重之间的相关系数。

II. 10 小鼠数据续集。为了一份生物作业，你量了同种的 12 只小鼠的尾长(厘米)及体重(克)。

- (a) 说明为什么你会预测：尾长和体重之间的相关系数会是正的。
- (b) 平均尾长算出来是 9.8 厘米。用英寸表示的平均尾长是多少?(1 英寸等于 2.54 厘米。)
- (c) 尾长和体重之间的相关系数计算出来是 $r=0.6$ 。如果你改用英寸而不用厘米量长度，则新的 r 值会是多少?

图 II. 3 里画的数据，是多种哺乳动物的平均脑重(克)对应平均体重(公斤)。有许多小型哺乳动物对应的点，在图的左下角堆叠在一起。习题 II. 11—II. 16 都是和这个散布图相关的问题。

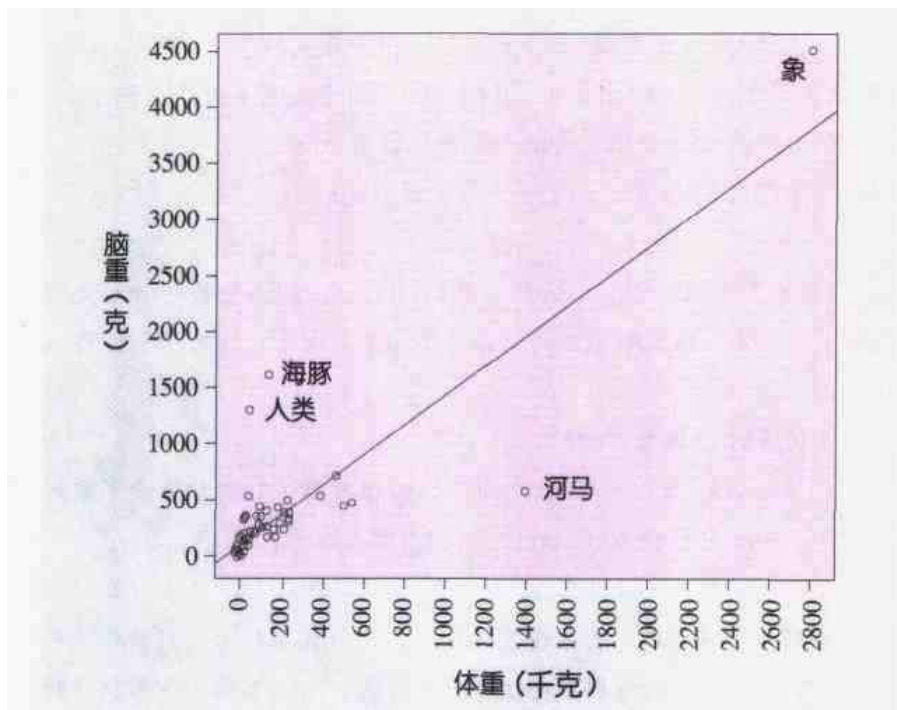


图 II. 3 96 种哺乳动物平均脑重(克)对应平均体重(千克)的散布图，对照习题 II. 11—II. 16

II. 11 海豚和河马。对应海豚和河马的点，在图 II. 3 中有特别标示。从图上读出这两种动物的体重及近似脑重。



II.12 海豚与河马。看到这个散布图的一个可能反应是：“海豚聪明，河马很笨。”是图的什么特点让我们有这种反应？

II.13 异常值。非洲象比这个数据组中任何其他动物都大得多，但仍然大致算是落在直线形态内。海豚、人类和河马就在直线形态之外。整组数据的脑重和体重相关系数是 $r = 0.86$ 。

- (a) 如果去掉象的那点，这个相关系数会变大、变小还是差不多？说明你的答案。
- (b) 如果同时去掉海豚、河马和人类，相关系数会变大、变小还是差不多？说明你的答案。

II.14 脑子和身体。体重和脑重之间的相关系数是 $r = 0.86$ 。用哺乳动物的体重来解释脑重，能解释多少？用一个数来回答这个问题，并简单说明这个数字提供了什么信息。

II.15 预测。图 II.3 散布图中的那条直线，是用体重预测脑重的最小二乘法回归直线。假设有人发现了一种躲在雨林中的新种哺乳动物，体重为 600 公斤。请预测这种哺乳动物的脑重。

II.16 斜率。图 II.3 散布图中的那条直线，是用体重预测脑重的最小二乘法回归直线。该直线的斜率是以下三个数字其中之一。请问哪个数字才是斜率？为什么？

- (a) $b = 0.5$
- (b) $b = 1.3$
- (c) $b = 3.2$

澳大利亚的包格斯先生提供了一组很特别的数据：每天早上在淋浴之前，他都会在他的淋浴间里称一称他的肥皂的重量，肥皂愈用重量就愈轻。数据就在表 II.3 里面（重量以克为单位）。我们看到其中有一天，包格斯先生忘了称肥皂。习题 II.17 到 II.19 的问题是针对这组肥皂数据问的。

II.17 散布图。把肥皂重量对应天数画图。整体形态是不是大致为直线？根据你的散布图，你觉得天数和重量之间的相关系数会符合以下哪一项：接近 1、大于 0 但不接近 1、接近 0、小于 0 但不接近 -1 或接近 -1？说明你的答案。



表 II.3 淋浴肥皂的重量

天数	重量	天数	重量	天数	重量
1	124	8	84	16	27
2	121	9	78	18	16
5	103	10	71	19	12
6	96	12	58	20	8
7	90	13	50	21	6

资料来源：包格斯先生。

II.18 回归。表 II.3 数据之最小二乘法回归直线方程式为

$$\text{重量} = 133.2 - 6.31 \times \text{天数}$$

- (a) 仔细说明斜率 $b = -6.31$ 对于肥皂重量减少的速率提供了什么信息？
- (b) 包格斯先生在第四天的时候没有量肥皂。用回归直线来推测那天肥皂的重量。
- (c) 把回归直线画在上一题的散布图上。

II.19 预测？用上---题的回归直线方程式来预测第 30 天的肥皂重量。为什么很容易知道你的答案毫无道理？用回归直线来预测第 30 天的重量有什么地方不对？

II.20 追随琼斯家的脚步。琼斯家在 1980 年时的住户收入是 30 000 美金，当时的 CPI(1982—1984 = 100) 是 82.4。2000 年初的 CPI 是 168.8。琼斯家在 2000 年必须有多少收入，才能拥有和 1980 年时同样的购买力？

II.21 拥有奔驰的代价。奔驰 190 在 1981 年要价 24 000 美元，当时的 CPI(1982—1984 = 100) 是 90.9。1999 年的平均 CPI 为 166.6。你要赚多少 1999 年的美元，才会有和 1981 年的 24 000 美元同样的购买力？

II.22 史坦威。一台史坦威演奏用钢琴在 1976 年的价格是 13 500 美元。类似的史坦威在 1999 年的价格是 79 900 美元。这台钢琴的真正价格是上涨还是下跌了？做计算来支持你的答案。



II. 23 金价：有些人建议投资者买黄金来“对付通货膨胀。”以下是1983年到1999年之间，每隔一年年底的1盎司黄金价格。画图来显示金价在这段期间的真正变化情形。如果投资黄金，是否能够保住它的真正价格，不受通货膨胀影响？

年度	1983	1985	1987	1989	1991	1993	1995	1997	1999
金价(美元)	385	329	486	403	354	391	392	368	295

II. 24 不在周日出院？加拿大的安大略省(Province of Ontario)针对国家医疗系统在该省的执行情况做了统计研究。图 II. 4 里的柱状图所根据的数据，来自对于安大略省社区医院病人住院及出院情形的研究。柱状图显示的是某两年期间内，一星期中每一天因心脏病开始住院以及出院的病人人数。

- (a) 说明一下为什么你会预期，因心脏病入院的病人人数，在一星期中的任一天应该都差不多。是否能从数据看出事实真的如此？
- (b) 描述一下病人出院的日子分布和入院的日子分布有何不同。你觉得为什么有这种差别？

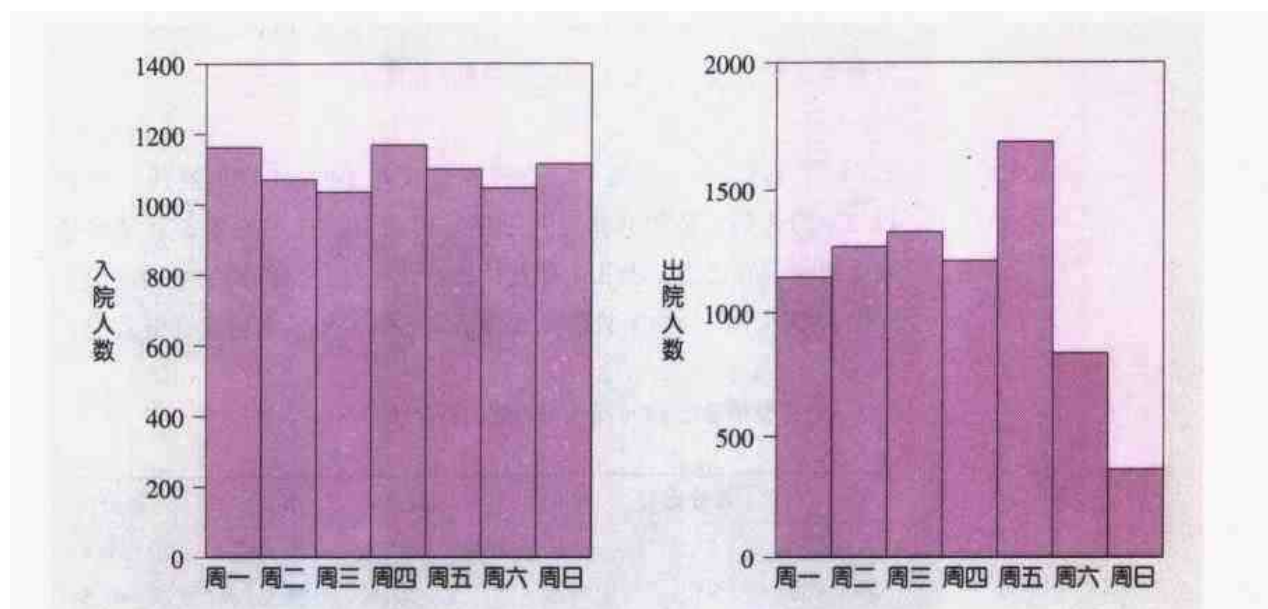


图 II. 4 加拿大安大略省医院的心脏病人，在一周中每一天入院及出院人数的柱状图，见习题 II. 24

II. 25 驾驶时间。穆尔教授住在大学城外数英里的地方，他每天都记录早上开车到大学去所花的时间。以下就是连续 42 个工作日的驾驶时间(分钟)，是一列一列照日子的顺序排的。



8.25 7.83 8.30 8.42 8.50 8.67 8.17 9.00 9.00 8.17 7.92
9.00 8.50 9.00 7.75 7.92 8.00 8.08 8.42 8.75 8.08 9.75
8.33 7.83 7.92 8.58 7.83 8.42 7.75 7.42 6.75 7.42 8.50
8.67 10.17 8.75 8.58 8.67 9.17 9.08 8.83 8.67

- (a) 替这些驾驶时间画个直方图。分布是大致对称、明显偏斜还是二者皆非?有没有明显的异常值?
- (b) 画一个驾驶时间的总图(把第1天到第42天标示在横轴上)。从图上看不出什么明显趋势,但是有一个特别短的驾驶时间,和两个特别长的时间。把这几个观测值在图上圈出来。

II.26 驾驶时间异常值。在上个习题中看到,墨尔教授开车上班的驾驶时间中有3个异常值,这3个都有合理解释。最短那次是感恩节的第二天(学校附近交通冷清)。较长的两次是因有交通事故和路上结冰而耽搁。把这3个观测值去掉。用计算机算出剩下39个时间的平均数 \bar{x} 和标准差 s , 来对正常驾驶时间做一个综合。

II.27 房价。2000年初独栋房子的中位售价是133 400美元。你认为平均售价会较高、差不多还是较低?为什么?

II.28 1996年美国选举。克林顿总统在1996年当选连任,得到49.2%的选票。他的共和党对手多尔得到40.7%的选票,其他候选人得到剩下的选票。表II.4有克林顿总统在每一州得到的全民选票百分比。用一个图,加上数值摘要以及简短的说明,来描述这组数据。

表II.4 克林顿总统1996年所得选票的百分比

州	百分比	州	百分比	州	百分比
亚拉巴马	43.2	路易斯安那	52.9	俄亥俄	47.4
阿拉斯加	33.3	缅因	51.6	俄克拉何马	40.5
阿里桑纳	46.5	马里兰	54.2	俄勒冈	47.2
阿肯色	53.7	马萨诸塞	61.5	宾州	49.2
加州	51.1	密歇根	51.7	罗得岛	59.7
科罗拉多	44.4	明尼苏达	51.1	南卡罗来纳	44.0
康涅狄格	52.8	密西西比	44.1	南达科他	43.0



(续表)

州	百分比	州	百分比	州	百分比
特拉华	51.8	密苏里	47.5	田纳西	48.0
佛罗里达	48.0	蒙大拿	41.2	德州	43.8
佐治亚	45.8	内布拉斯加	35.0	犹他	33.3
夏威夷	56.9	内华达	43.9	佛蒙特	53.4
爱达荷	33.6	新罕布什尔	49.3	弗吉尼亚	45.2
伊利诺伊	54.3	新泽西	53.7	华盛顿	49.8
印地安纳	41.6	新墨西哥	49.2	西弗吉尼亚	51.5
艾奥瓦	50.3	纽约	59.4	威斯康星	48.8
堪萨斯	36.1	北卡罗来纳	44.4	怀俄明	36.8
肯塔基	45.8	北达科他	40.1		

资料来源：美国联邦选举委员会。

II. 29 投资统计。乔参加的退休金计划将钱经由“指数基金”投资于股市，指数基金跟随着用标准普尔 500 指数来度量的大盘而起舞。乔想要买不太和大盘连动的共同基金。他看到报道，富达科技基金的每月获利和标准普尔 500 指数之间的相关系数是 $r=0.77$ ，而富达房地产基金和 500 指数的相关系数是 $r=0.37$ 。

(a) 这两种基金当中哪一种的获利和大盘的关系比较密切，你怎么知道的？

(b) 上面这些资料有没有提供乔任何关于哪个基金获利较高的信息？



第二部分 报告作业

报告作业是比较长的习题，需要搜集信息或制作数据，而且重点是要把做出的结果用短文来说明，这里很多题目适合由一组学生共同来做。

作业 1. 新闻界使用的统计图。不论好的图还是坏的图充斥于各种新闻媒体。有些出版品，例如《今日美国》，更是经常用图来呈现数据。从报纸和杂志上(不要用广告)搜集几个图(至少 5 个)。其中要包括你认为优良的图和你认为很差或者误导的图。把你搜集的图当做例子，写一篇短文讨论媒体所用的图清楚与否、精确程度以及是否具有吸引力。

作业 2. 回归自选题。自己选择两个你认为与大致直线相关的数量变量。针对这个两个变量搜集数据并做统计分析：画散布图，找出相关系数及回归直线(用计算机或统计软件)，并把直线图在散布图上。然后把结果写份报告。以下是一些可以考虑的变量例子。

- (a) 一群人的身高和两臂伸展的张距。
- (b) 一群人的身高和步距。
- (c) 数个不同品牌，不同大小瓶装洗发精的每盎司价格及瓶子容量(单位：盎司)。

作业 3. 高中辍学生。写一篇关于美国高中辍学生的真相报告。以下是你可以谈的问题的一些例子：哪一州的成人未读完高中的百分比最高？辍学生的收入和就业状况与其他成人比起来如何？没有读完高中的百分比，是不是黑人和西班牙语系的人比白人要高？

《美国统计精粹》里可以找到许多数据。在教育(education)项下找高中辍学生(high school dropouts)。也许你也有兴趣看看《美国统计精粹》里其他部分，有关收入及其他变量在不同教育程度下的数据。

作业 4. 相关不代表因果。用“相关不代表因果关系”为主题，写一篇很漂亮且引人注意的文章。用像第 15 章的例 6 及习题 15.21 那样针对某件事情但半开玩笑的例子，或者像下面这样的例子：长头发



和女性之间有相关关系，但是把一位女性的头发剪短并不会把她变成男性。说明要清楚，但是不要太专业了。当做你的读者是高中生来写这篇文章。

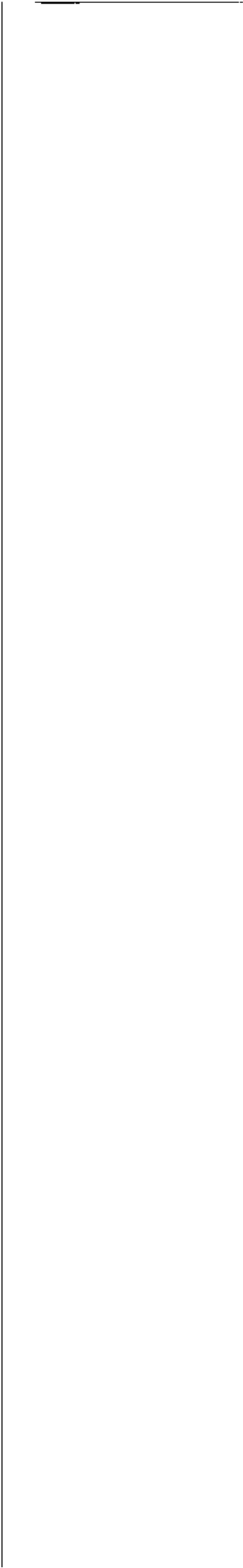
作业 5. 军事支出。以下数据摘自《美国统计精粹》，是 1940—1999 年间的会计年度美国在国防方面的支出。你也许有兴趣从最新的卷期当中找出最近的资料。单位是百万美元。

年度	1940	1945	1950	1955	1960	1965	1970
军事支出	1 660	82 965	13 724	42 729	48 130	50 620	81 692
年度	1975	1980	1985	1990	1995	1999	
军事支出	86 509	133 995	252 748	299 331	272 066	276 730	

写一篇短文，描述在二次大战之前到冷战结束 10 年后这段期间，军事支出的真正改变情形。要做必要的计算，并简短描述军事支出和以下所列这段期间的主要军事事件之间的关系：第二次世界大战（1941—1945）、朝鲜战争（1950—1953）、越战（大约 1964—1975）、以及 1989 年柏林围墙倒下后冷战结束。

作业 6. 你的脉搏。你的“静止时脉搏”（resting pulse rate）是多少？当然即使你在静止时量自己的脉搏，每天量的也不见得一样，而同一大不同时间量的也不见得相同。量一下你静止时的脉搏，同一天内至少量 6 次（时间隔开些）而且至少要量 4 天。写一篇报告，内容必须包括你是怎样量脉搏的，以及你对数据做的分析。根据你的数据，当别人问你静止时的脉搏多少，你会怎么回答？（如果有好几位学生做了这题作业，你们也可以讨论一群人的脉搏变异情况。）

作业 7. 铜板上的日期。许多铜板上都刻有铸造年份。对以下每一个面值的铜板都至少搜集 50 个铜板的资料：一分、五分、一角及二角五分。描述一下现在仍在流通的铜板的年份分布，要包括图以及数值描述。不同面值的铜板间有差别吗？有没有发现异常值？



第三部分

机遇

“如果我有当国王的机遇，那么我就有戴上皇冠的命。”麦克白在莎士比亚剧中这样说。机遇(chance, 也做可能性或机会)的确会作弄我们每个人，而我们却没多少能耐可以了解或者控制它。不过有的时候机遇被驯服了。掷骰子、简单随机样本，甚至得自遗传的眼珠颜色或者血型，代表已经安分下来的机遇，是我们可以了解并且控制的。不像麦克白或者我们的一生，骰子是可以一掷再掷的。结果是由机遇决定，但是在多次重复之后，会有某种模式(pattern, 也译做形态)出现。而因为我们可以描述它的模式，也使得机遇不再神秘莫测了。

不论是几何里的圆和三角形，还是行星的运行，我们人类都用数学来描述它们规则的模式。当我们可以不断重复某种机遇现象，使得机遇变得可掌握时，我们会利用数学来了解这个机遇现象的规律模式。机遇的数学叫做概率(probability)。概率就是这一部分的主题，然而我们会少讲数学而着重实验和思考。

第 17 章

考虑可能性

可能性有多大？

1986 年 1 月 28 日那一天，挑战者号航天飞机在发射后不久，就在世人眼前爆炸了。总统特别委员会开始调查，参与该项发射计划的人估计像这样发射失败的机会有多大？部分直接参与的工程师说，大约是百分之一的机会。管理部门说，大概 10 万次才会发生一次。在听到后面这项估计之后，委员会成员之一的物理学家费曼 (Richard Feynman, 1918—1988) 就问：“你们的意思是说，如果连续 300 年每天发射一次火箭，你们预期只会失败一次？”费曼的心算很棒：假如不考虑闰年的话，300 年等于 109 500 天。

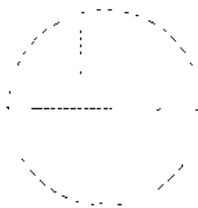
费曼在这次简短的对话中做了两件很重要的事。管理部门很明显



只是在猜(好吧,他们自以为是在根据资料对风险做判断,可是结果是一样的)。也就是说,他们在用可能性的语言,来表达他们的个人意见或判断。费曼把这个模糊的个人意见、当试改用较具体的意象来表达,也就是同一件事重复做许多次的概念:如果我们发射了非常多的航天飞机,那失败的频率大概会是多少?这点听起来应该很耳熟吧,因为我们对抽样就是这样考虑的:“如果我们从同一总体抽取许多样本,总体的真正值会有95%的时候落在误差界限之内。”

费曼用的第二个技巧,是通过和真实生活相联系,而让人对于尝试某件事100 000次的意义更容易了解,他把这个转换成每天试一次,共试300年。但这样子还是不容易弄清楚。我们的脑袋对于很大的数字不大反应得过来,对很小的概率也是一样,比如8 000万次中有一次赢得乐透彩的机会,或者搭乘700万次飞机时会有死于坠机的机会。

可能性是不大容易掌握的题目。我们会追随费曼,从“如果我们试很多次,会发生什么情况”开始,然后才会试着考虑,怎样用可能性的语言来表达个人意见。我们也会先讨论像掷铜板时有二分之一机会得到正面这样的例子,然后才来考虑乐透彩。



概率的概念

即使美式足球规则也都认为,掷铜板可以避免偏袒。抽样调查挑选受访对象时,或者医学试验将病人分配到处理组或安慰剂组时,如果有偏袒,就像美式足球比赛开始时决定球先给哪一队时有所偏袒一样,都是不能接受的。这就是为什么统计学家建议使用随机样本及随机化实验,这些只是掷铜板的花哨版本。如果我们仔细观察掷铜板或者随机样本的结果,一件重要事实就会浮现:短期机遇现象无法预测,但是长期下来,会呈现有规则且可预测的模式。

掷一个铜板或者选择一个简单随机样本,都无法在事前预测结果;因为你如果重复掷铜板或选样本,结果就会次次不同。但是还是可以在结果里面看到某种规则模式,而只有在重复许多次以后,这个模式才会清楚浮现。这个了不起的事实,就是概率概念的基础。



例1 掷铜板

当你掷铜板的时候，结果只有两种可能，正面或者反面。图 17.1 显示掷铜板 1 000 次的结果。对应从 1—1 000 次的每一掷，我都将掷出正面的比例阵在图上。每一次掷出正面，所以正面比例的第一个值是 1。第二次掷出反面，所以在两掷之后，正面比例降为 0.5。再接下去的三次是反面之后两个正面，所以掷五次之后的正面比例是 $3/5$ ，即 0.6。

刚开始的时候，正面比例变化很大，但是掷的次数越多，就慢慢稳定下来。这个比例到最后会靠近 0.5，而且会一直维持在 0.5 附近。我们把这个 0.5 叫做得到正面的概率。0.5 这个概率在图上是以水平线表示出来的。

在统计里的“随机”(random)，并不是“偶然”(haphazard)的同义字，而是在描述某种长期下来才会出现的规则。我们在每一天的生活经验里，都会碰到随机不可预测的那一面。但我们很少有机会能重复观察同一个随机现象许多次，而且次数多到能够看出概率所描述的长期规则模式。你可以在图 17.1 中看到规则出现，从长期来看，掷

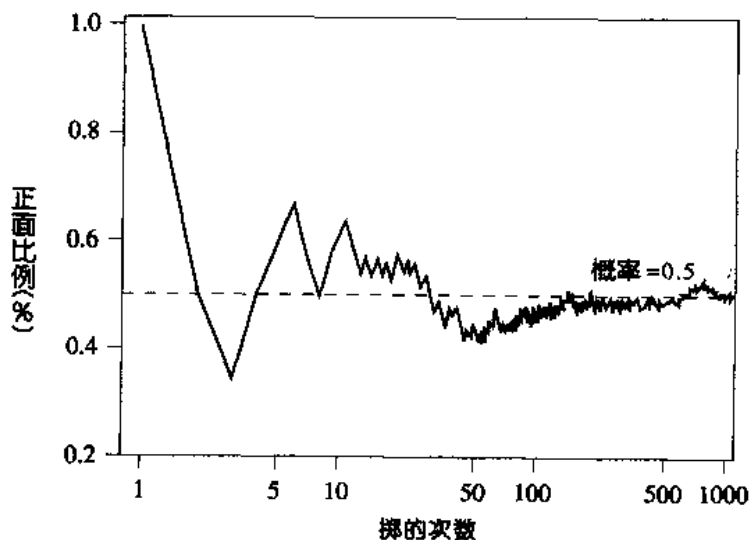


图 17.1 掷铜板许多次，掷出的正面比例随着我们掷的次数而改变，但最终会非常接近 0.5。这就是我们说“正面概率是一半”的意思



出正面的比例是 0.5，这是概率的直觉概念。概率 0.5 代表“试验很多很多次时，有一半时候会发生”。

我们可能仅仅因为铜板只有两面，就会猜正面概率是 0.5 了。但是婴儿的性别也只有两种可能，概率却不一样——男婴的概率差不多是 0.51，而不是 0.50。这个概率概念是根据经验法则而来，也就是说，是根据数据得来而不是根据理论。概率描述在多次试验中会发生什么情况，而我们必须真的观察许多次掷铜板的结果或许多的婴儿，才能够掌握这个概率。说到掷铜板，有些勤快的人还真的掷过成千上万次呢。

例 2 掷铜板的人

法国自然主义者布方伯爵(Count Buffon, 1707—1788)把铜板掷了 4 040 次。结果：2 048 个正面，或者说正面比例是 $2\,048/4\,040 = 0.5069$ 。

大约 1900 年时，英国统计学家皮尔逊(Karl Pearson, 1857—1936)很神勇地掷一个铜板 24 000 次。结果：12 012 次正面，比例 0.5005。

南非数学家柯瑞屈(John Kerrich)在第二次世界大战被德国人关在牢里的时候，掷了铜板 10 000 次。结果：5 067 次正面，比例 0.5067。

• 随机及概率

如果一个现象的个别结果无法预知，然而在多次重复之后，其结果会出现有规则的分布，则我们称该现象为随机的。

一个随机现象任一结果的概率是在 0—1 之间的一个数字，该数字描述在重复许多次的情形下，其结果应会出现的比例。

概率为 0 的结果从来都不会发生，而概率为 1 的结果则每重复一次就发生一次。概率为 $1/2$ 的结果，在一长串的试验当中，大约有一半时间会发生。当然我们永远没办法确实观察出一个概率。比如说，铜板不管掷了多少次都可以再掷。而数学的概率是一种理想化的



描述，根据的是推想在一串无休无止的试验当中会发生的情况。

我们没有要在这里做深入的研究。随机现象的存在，只不过是我们在观察这世界所得到的事实。概率也不过是用来描述随机现象长期规律性的语言。掷一次铜板的结果、放射源发射出粒子的间隔时间以及实验室老鼠生的下一胎小老鼠的性别都是随机的。随机样本或者随机化实验的结果也一样是随机的。一大群人的行为，也常常和掷多次铜板或取多个随机样本的结果差不多随机。举例来说，人寿保险根据的就是在一大群人里面，死亡是随机发生的这项事实。

例 3 死亡的概率

我们没法预测特定的人是不是明年会死。但是如果我们观察好几百万人，死亡就是随机的了。国家卫生统计中心宣布，20—24 岁的男性当中，在任一年中会死的比例差不多是 0.001 5。这是一个年轻男人明年会死的概率。对于同年龄层的女性，死亡概率大约是 0.000 5。

如果一个保险公司卖出许多人寿保险给年龄在 20—24 岁之间的人，公司会知道：卖给男性的保险明年大约有 0.15% 要理赔，卖给女性的大约有 0.05% 要理赔。而因为男性理赔的比例比较高些，所以保险费会收得多一点。

机遇的古代史

在玩需要多次重复的机遇游戏，如掷骰子、发洗好的牌、转轮盘等时，我们最容易注意到随机现象。类似这些游戏的机遇装置，曾在远古时代用来找寻神的旨意。西方古时候最常用来随机化的方法是“掷骨头”，就是掷好几块距骨(图 17.2)，距骨是相当规则的实心骨头，取自动物的脚跟。掷了之后，等距骨静止，四面中的其中一面会朝上(其他两面是圆形的)。用陶土或骨头做的立方体骰子后来才有，但即使是骰子，在公元前两千年之前都已存在。跟占卜术比起



来，用掷骨头或掷骰子来赌博几乎算是较近代的发展。大约公元前三百年前的时候，还没有这种“堕落行为”的明确记录。赌博在罗马时代大为风行，后来在基督教的不认同之下暂时消退(和占卜一起)。

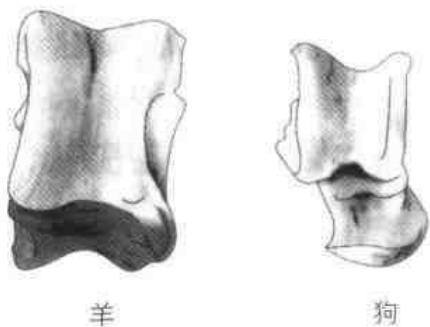


图 17.2 动物的距骨(邱意惠绘)

有史以来，诸如距骨这样的机遇装置就被使用了。但是，古代那些伟大的数学家中，没有一个人研究过掷骨头或掷骰子很多次得到的规律模式。也许是因为距骨以及大部分的古代骰子形状不够规则，使得每一个的结果都有不同的模式。或者原因是更深层的，因为传统上不情愿做有系统的实验。

职业赌徒不像哲学家或数学家那么抑制自己，他们注意到掷骰子或发牌的结果有规则模式，并试着调整赌注来增加赢钱机会。“我该怎么赌？”这个问题，就是概率理论的起源。对于随机的有系统研究，是从(我有一点过度简化)17世纪时法国赌徒请法国数学家帮忙算出机遇游戏的“公平”赌注时开始。“概率理论”也就是随机的数学研究，是17世纪时从费马(Pierre de Fermat, 1601—1665)及帕斯卡(Blaise Pascal, 1623—1662)开始的，到20世纪统计学家接手的时候，概率理论已经发展得很完善。

关于机遇结果的神话

概率的观念似乎很直截了当。它对“如果我们这样做很多次，会发生什么情况？”这个问题，提供了答案。但事实上不论是随机现象的“表现”，还是概率观念，都有很微妙的地方。我们不断地会遭遇机遇结果，而心理学家告诉我们，我们处理得不高明。

短期规律性的神话。概率的概念是，随机现象长期来说是有规则

上帝掷骰子吗？

世界上很少有事情是百分之百的随机，以致于不管我们有多少信息，都没法子预测结果。比如说，理论上我们可以把物理定律应用在掷铜板上面，来计算会掷出正面还是反面。但是在每个个别的原子内部，随机性的确规范了事件的发生状况。爱因斯坦(Albert Einstein, 1879—1955)不大喜欢新量子论的这项说法。

“上帝可没有在和宇宙玩骰子”，这位伟大的科学家说。但80年之后看来，爱因斯坦显然错了。



的。不幸的是，我们对于随机的直觉却是说，随机现象应该在短期就有规则。当规则没出现时，我们就会寻求解释，而不把它当作是机遇变异。对了解机遇而言，我们的直觉会给我们很差的指引。

例 4 什么看起来像随机的？

把一个铜板掷 6 次，并且把每次是正面或反面记录下来。以下哪个结果比较可能发生？

正反正反反正 反反反正正正

几乎每个人都说“正反正反反正”比较容易发生，因为反反反正正正“看起来不随机”。事实上，两者发生的机会一样大。正面和反面机会均等的意思只是说：掷了很长一串的结果里，应该大约有一半是正面；可没有说只掷很少次时正反就应该差不多是间隔发生。铜板没有记忆，又不知道前几次掷了什么，没法试着制造出一串平衡的结果。

掷 6 次铜板得到反反反正正正这样的结果看起来不寻常，是因为有连续 3 个反面和连续 3 个正面。连续出现同样的东西好像在直觉上“不随机”，但实际上常发生。以下这个例子比掷铜板还要令人印象深刻。

例 5 篮球赛中“手风正顺”的球员

认为一连串同样结果必定不是由于机遇产生，这个信念会改变人的行为。如果一个篮球运动员连续几球都投进，球迷和他的队友就会相信他“手风正顺”，下一球很可能又投进。



这是不正确的。严谨的研究显示出：“在篮球赛中球员连续进球或连续不进球发生的频率，与每一球和前一球互相独立情况下预期的频率比起来，前者并不会比较频繁。球员的表现是一贯的，不是一阵子好、一阵子坏。如果一个球员的长期命中率是一半，那么他投中或不中的情况就像掷铜板一样，那就是说，连续进球或连续不进球发生的机会，比我们直觉以为的要大。

意外机遇的神话。茱莉在伦敦过暑假。有一天，在维多利亚及亚伯特博物馆的三楼，她遇见了大学里认识的普通朋友吉姆。“真难得！也许我们命中注定会相遇。”

恐怕未必见得。茱莉要在那一天遇到这个朋友的机会当然很小，但是茱莉待在伦敦的这个暑假中，会碰到某个认识的人的机会可不是很小。毕竟，一个普通的成人认识的人约有1 500个。当不寻常的事发生，我们会回头想想并且说“怎么会有这么巧？”如果其他1 500件不容易发生的事当中的任何一件发生了，我们也会有同样的反应。在以下例子当中，我们还可以真的算出概率。

降雨概率是……

工作了一个礼拜，然后周末却下起了雨。我们觉得天气老跟我们作对的这种想法，可不可能背后真的有统计事实支持？至少美国东岸的答案是肯定的。让我们回到1946年，似乎周日下的雨比周一多22%。可能的解释是，工作日在路上跑的汽车和卡车所造成的空气污染，有助于形成雨滴，而它作用所需的时间又恰好让雨下在周末。

例6 中两次头彩

1986年时，亚当斯(Evelyn Marie Adams)第二度赢得新泽西州彩券，前一次亚当斯赢到了累积奖金390万美元，这次又赢了150万美元。《纽约时报》(1968年2月14日)宣称：同一个人赢两次大奖的机会，差不多是每170亿次中有一次。两星期后，《纽约时报》刊登了两位统计学家的来信，说这是胡说八道。亚当斯在一生中赢两次大奖的机会诚然很小，但是几乎可以确定：在美国几百万经常买彩券的人当中，会有人赢两次累积奖金。两位统计学



家估计：7年内再有人赢到两次大奖的机会是一半一半。果不其然，在1988年5月，汉弗莱斯(Robert Humphries)赢得了他的第二个宾州彩券累积奖金(总计680万美元)。

不寻常的事件，尤其是令人悲痛的那些，会让人很想要找出一些道理，也就是造成该结果的“因”。在这里我们给以前对于因果关系的讨论加以补充：有时就只是机遇巧合罢了。

例7 癌症群

1984年时，在马萨诸塞州兰道夫某一个区域，250户中有67个癌症个案。这样集中的癌症个案有点不寻常，住户担心：附近一个化学工厂的排放物污染了用水，因此而致癌。

1979年时，马萨诸塞州沃本镇(Woburn)使用的八口井中，发现有两口井遭有机化学物污染。感到忧虑的镇民开始计算癌症病例。在1964—1983年之间，沃本镇登记在案的儿童白血病共有20件。对这种较稀有的病来说，这样的病例数字并不寻常。镇民相信井水造成了白血病，开始控告两家应对污染负责的公司。

癌症是普遍的疾病，在美国是超过23%的死亡原因。有时癌症病例会在邻近区域内密集发生，这不算稀奇：总有“某个”地方因为巧合就有多个癌症病例发生。可是当癌症群(cancer cluster)发生在“我们”邻近的区域时，我们就会往坏的方向想，而想找某个人来责怪。美国各州政府每年都会接到几十个老百姓的电话，担心他们住的地区“有太多癌症”。但是诚如美国国家癌症研究所所说的：“大部分的癌症集中情况，只不过是巧合罢了。”

对马萨诸塞州的两个癌症群，哈佛大学公共卫生学院的统计学家都进行过调查。调查人员试着取得曾在问题发生期间住在该区域的每



一个人的完整资料，并估计每一个人和可疑饮水的接触程度。调查人员也试图取得其他可能致癌因素的相关资料，例如吸烟与否，以及在工作时是否会接触有毒物质。最后的判决是：对兰道夫的癌症群而言，可能的解释是凑巧；但是却有证据显示，在沃本镇的两口井取得的饮水和得儿童白血病之间有相关关系。

平均数定律的神话。有一次在拉斯维加斯开大会时，我在赌场漫步，眼看着钱从桌上消失，落入赌桌下庄家的盒子里。你在赌场里会看到有趣的人类行为。当掷骰子的人连续赢了几把的时候，有些赌徒会认为她“手风正顺”，打赌她还会继续赢，其他人却说，根据“平均数定律”（law of averages），她应该要输了，这样输赢才能平衡。笃信平均数定律的人认为，如果你掷铜板6次而得到反反反反反反，下一次掷一定是得正面的概率比较大。长期来说正面的确应该占一半。所谓的神话是指：认为像连出了6个反面这样的不平衡状况，会在下一次的結果中得到补救。

铜板和骰子没有记忆。铜板不知道前6次的结果是反面，不能在下次掷的时候想办法得个正面来平衡一下。当然，长期下来真的会达到平衡。在掷了10 000次以后，头6次的结果就无足轻重了，但不是被“补救”，而是被后来掷的9 994次的结果淹没了。

例8 我们要儿子

相信这个假的“平均数定律”，有可能导致近乎灾难的结果。几年前，“亲爱的艾比”这个提供建议的专栏，刊登出了一个心烦意乱的母亲的信，这母亲一连生了8个女儿。似乎原来她和她先生只准备要4个孩子的，可是当4个都是女孩的时候，他们就再试一次，且一试再试。在连续7个女儿之后，连她的医生都向她保证：“根据平均数定律，成功和失败的机会是100比1。”不幸的是，这对夫妇来说，生孩子就和掷铜板一样。连续8个女孩发生的机会很小，但是在已经有7个女孩子之后，下一个还是女孩的机会并不小——而且也的确发生了。



个人概率

乔坐在那儿瞪着他的啤酒，他心爱的棒球队，芝加哥小熊队，刚刚又输了一场球。小熊队拥有一些很好的年轻球员，所以我们来问问乔：“明年小熊队参与世界大赛的机会会有多大？”乔的眼睛亮起来了，“噢，大概 10%。”他说。

是乔把小熊队能打进世界大赛的概率定成 0.10 吗？下一年的比赛战果当然是没法预知的，但若我们考虑重复许多次会发生什么情形，又不大合理。明年的棒球季只会发生一次，而且在球员、天气和其他许多方面都会和其他球季不同，我们问的问题答案似乎很清楚：如果概率度量的是“假如我们重复许多次，会发生什么状况，”则乔说的 0.10 根本不是概率。概率是根据同一个随机现象重复许多次所得数据而来的。乔给我们的不是这个，而是他的个人判断。

可是常常当我们在用“概率”这个词的时候，也包括了我们对于某个事件发生的可能性的个人判断。我们还根据这些判断做出决定——我们搭公共汽车进城，因为觉得能找到停车位的概率很低。连更重要的决定都会把对“机会有多大”的判断列入考虑。要决定是否要建新厂的公司现在就必须判断，当三年后新厂盖好时，消费者对该公司产品有大量需求的机会会有多大。许多公司把他们对“机会有多大”的判断，用数字表示并把它当成概率，还用来计算。三年后有大量需求，就像小熊队赢得明年的参赛权一样，都是“只此一次”的事件，没法适用“重复许多次”这种思维方式。还不止这样呢，公司的每个高级主管可能给的概率都不一样，反映出他们每个人的判断都不同。因此我们需要另外一种概率——个人概率(personal probability)。

• 个人概率

一个事件的个人概率(personal probability)是 0—1 之间的一个数字，代表个人对于该事件发生机会有多大的判断。

个人概率有个人优势，就是不限于“能重复的情境”。这种概率很有用，因为我们根据它做决定：“我相信海盗队会赢超级杯的概率



是 0.75，所以我要去赌这场比赛。”要记住个人概率和“重复许多次的比例”这种概率是不同种类的。前者只代表个人意见，无所谓对还是错。

即使在可以“多次重复”的情境下，这点还是对的。如果克瑞的直觉告诉他，下一次掷铜板出现正面的概率是 0.7，这就是克瑞的想法，就是如此而已。若把铜板掷很多次，可能显示出正面的比例很接近 0.5，但那是另一回事。规定个人对于一次试验结果的信心，必须和试很多次的结果一样，是没有道理的。我特别强调这点，因为常常有人认为“个人概率”和“试很多次会发生什么状况”不过是同一个观念的两种不同解释，但事实上这两个观念的差别很大。

为什么对于个人意见我们还要用“概率”这样的字眼呢？有两个很好的理由，首先，如果我们知道试验很多次的结果，则我们通常也的确会根据这些数据来做个人判断。布方伯爵、皮尔逊及柯瑞屈掷铜板的结果(例 2)，或者可能由于我们自己的经验，让我们相信掷铜板许多次的话，正面出现的次数很接近一半。当我们说这次掷铜板，出现正面的概率是 $1/2$ 时，我们是在把根据掷很多次会发生的结果，而得到的正面概率的量度，应用在掷一次的状况上面。其次，个人概率或长期比例的概率，都遵循同样的数学规则，例如，两种概率都是在 0—1 之间的数字。这些对我们来说，并不如数学家那样重要，不过我们在下一章还是会介绍一些概率规则。而这些规则对两种概率都适用。

胜算如何？

赌博者通常用胜算(odds)而不是用概率来表达机会。不利于某事件发生的胜算是 A 对 B，代表该事件发生的概率是 $B/(A+B)$ 。所以“胜算为 5 对 1”是“概率为 $1/6$ ”的另一种说法。概率必定介于 0 与 1 之间，但胜算的范围可以从 0 到无限大。虽然胜算主要用在赌博，我们还是可借助它把很小的概率表达得更清楚。“胜算是 999 对 1”可能比“概率是 0.001”更容易理解。

概率及风险

一旦我们知道，“对于机会多大的个人判断”和“重复许多次会发生什么情况”是不同的概念，就可以了解为什么一般大众和专家，对于什么时候风险很大、什么时候不会有很人风险的意见会大不相同。专家是用根据数据算出的概率，来描述遇上某个不受欢迎事件的风险；然而个人或者社会却似乎对数据置之不理。我们为一些几乎永远不会发生的事担心，却对某些更有机会发生的事毫不在意。

为什么我们把石棉的风险看得比驾驶的风险重得多？为什么我们对一些很难碰上的威胁，像龙卷风和恐怖分子，担忧的程度超过担忧得心脏病？



例 9 学校里的石棉

高度暴露于石棉是危险的。但低度暴露的风险却很低,例如,学校的暖气管周围隔热材料中有石棉,学校里的老师和学生的风险很低。一位老师如果在一个有典型的石棉含量的学校里工作三十年,他会因石棉而得癌症的概率差不多是 $15/1\,000\,000$ 。开车的人一辈子当中,会死于车祸的概率大约是 $15\,000/1\,000\,000$ 。也就是说,经常开车的风险是在有石棉的学校里教书的风险的 1 000 倍。

有风险并没让我们停止开车。但是风险小得多的石棉却引发了大规模的清除运动,美国联邦政府还要求每个学校必须检查石棉并公布结果。

- 比较起来,当风险似乎在我们掌握之中时,我们会比在不能控制它时更觉得安全。我们开车时可以掌握情况(或者自以为如此),但对于来自石棉、龙卷风或恐怖份子的风险,我们完全不能控制。
- 要了解非常小的概率有点困难。 $0.000\,015$ 和 0.015 的概率都很小,我们的直觉不能分辨出其间的差别。心理学家曾指出,我们通常会将很低的风险高估,而将较高的风险低估。也许,这就是我们对概率运作的直觉的一个普遍弱点。
- 像学校里的石棉这一类风险的概率,不像掷铜板的概率那样确定,必须由专家经由复杂的统计研究来估计。也许最安全的做法,是怀疑这些专家可能低估了风险程序。

我们对于风险的反应,也不光是由概率决定的,即使我们的个人概率已经比专家根据数据算出的概率要高时,仍然如此。我们会受自己的心太以及社会规范的影响。诚如一位作者曾经说的:“即使撞车的风险远高于家里出事的风险,但我们就算只开车出去办 10 分钟的事,也很少有人会把婴儿独自留在家睡觉。”



网络寻奇

要掌握概率概念的最好方法之一，就是能眼看着一个事件发生的比例，随着试验次数的增加而逐渐在该事件的概率附近稳定下来。电脑模拟可以做到这点。去《统计学的世界》原文版网站：www.whfreeman.com/sec，点击概率小程序中的 What Is Probability。

癌症群引起大众高度关切，以至于美国国家癌症研究所特别为这个主题专辟了一个网页：http://cis.nci.nih.gov/fact/3_58.htm。



本章重点摘要

世界上有些事是**随机**的，不管天然的和人工的都有。也就是说，虽然这些事每一次试验的个别结果无法预测，但在极多次重复之下结果会出现明显的模式。我们用概率来描述随机现象的长期规律性。一个事件的概率，是重复许多许多次之后，该事件发生的比例。概率是 0(从不发生)—1(必定发生)之间的数字。我强调这种概率，是因为它是根据数据得来的。

概率只描述长期下来发生什么事。像掷铜板或投篮之类随机现象的短期表现常常看来不随机，是因为次数不够多，所以看不到只有在极多次重复时才会出现的规则。

个人概率代表一个人对某件事发生机会的个人判断。个人概率也是在 0—1 之间的数字。不同的人可能提出不同的个人概率，而且个人概率不见得是根据类似情形下发生比例的数据而来。



第 17 章 习题

17.1 旋转铜板。把一个铜板立在一个硬的表面上，用食指压住，然后用另一只食指去弹它，让它旋转并等它倒下。转 50 次，估计铜板正面向上的概率。

17.2 坠落的铜板。你可能认为因为铜板有两面，所以掷铜板时得到正面向上的概率理当约等于 $1/2$ 。这种想法不见得一定对。上一题要求你旋转铜板而不是掷它，这样正面概率就不同了。再试另一种做法：把一个铜板竖在硬的平面的边缘。用手拍击该平面让铜板掉下去。掉落后正面朝上的概率是多少？试做至少 50 次来估计正面概率。

17.3 随机数字(random digit)。表 A 里的随机数字表是由一种随机装置产生，表里面的任一个数字是 0 的概率为 0.1。表里面的头 200 个数字中，0 占的比例是多少？这个比例是真正概率的估计值，是重复 200 次所得到的，而这个例子中的真正概率是 0.1。

17.4 掷多少次才得正面？我们掷铜板的时候，根据经验正面概率（长期比例）大约 $1/2$ 。假设现在我们来掷铜板，掷到一出现正面就停。第一个正面会出现在奇数次（1、3、5 等等）的概率是多少？为了找到这个概率，重复这项实验 50 次，把每一回合实验当中需要掷几次才出现正面的次数记录下来。

(a) 根据你的实验结果，估计第一次掷就掷出正面的概率。我们会预期这个概率是多少？

(b) 用你得出的结果，来估计第一次正面是出现在奇数次的概率。

17.5 掷图钉。把一颗图钉在硬的表面上掷 100 次。有多少次图钉的尖端是朝上的？图钉“着陆”时会尖端朝上的近似概率是多少？

17.6 三条。你在一本有关麻将牌的书中读到，发 5 张牌时会得到三条的概率是 $1/50$ 。用简单的语言说明这句话的意思。

17.7 从语言到概率。概率是一个事件发生机会有多大的一种量



度。把以下列出的概率和有关可能性的叙述做一个配对。(以概率为可能性的量度,通常比言语叙述要更确实。)

0 0.01 0.3 0.6 0.99 1

- (a) 这个事件不可能,它永远不会发生。
- (b) 这是个必定发生的事件,每次试验它都会发生。
- (c) 这个事件机会很小,不过在一长串的试验中偶尔会发生。
- (d) 这个事件发生的机会比不发生的机会大。

17.8 打赢棒球赛。1969—1989 年这段期间,美国职业棒球两大联盟的冠军队,球季中在主场出赛的胜率是 63%。赛季结束时,两大联盟的冠军队碰头,角逐世界大赛。你会不会应用之前的结果,而认为主场球队会赢得世界大赛的概率是 0.63?说明你的答案。

17.9 你开车会出意外吗?任意选出的一个驾驶员,明年会发生事故的概率是 0.2。这个数字是根据几百万开车的人的事故比例得来的。这里的“事故”不是只指公路上的车祸,也包括像你在自己家车道撞凹挡泥板这类事。

- (a) 你觉得明年你会碰上事故的概率是多少?这是个人概率
- (b) 提出一些理由说明,为什么对于预测你会有事故的概率来说,你的个人概率,可能还比任选一个驾驶人所得到的概率要准确些。
- (c) 差不多每个人都认为,自己会出事故的个人概率,比任选驾驶人的概率要低。你想这是为什么?

17.10 婚姻状况。任选一位 40 岁的女性,她已离婚的概率是 0.16。这个概率是长期以来,几百万个 40 岁女姓的离婚比例。我们假设 0.16 这个概率在接下来的 20 年都不会变动,而碧姬现在 20 岁,还没结婚。

- (a) 碧姬认为她自己在 40 岁时会已经离婚的概率大约是 5%。说明为什么这是个人概率。
- (b) 提出一些理由,为什么碧姬的个人概率可能和所有 40 岁女姓的离婚比例不同。
- (c) 你是政府官员,负责研究社会安全保险制度对中年离婚女性的影响。你只在意 0.16 这个概率,而对于任何人的个人概率都不感兴趣。为什么?



17.11 选择个人概率还是数据。举个例子，是你会应用由多次试验的数据所得的长期比例当做概率的情况。举例说明你宁愿用个人概率的情况。

17.12 个人概率？当数据很少的时候，我们通常只能依靠个人概率。在挑战者号发生灾难之前，总共只发射过 24 次航天飞机，而且全部都发射成功。航天飞机计划的管理层认为，发射航天飞机失败的概率只有 0.000 001。提出一些理由来说明，为什么这样的估计多半太过乐观。

17.13 个人随机数字。找几个朋友（至少 10 个），请他们每人“随机”给一个 4 位数。这些数字当中有几次以 1 或 2 开头？几个以 8 或 9 开头？（有很强的证据显示，一般人倾向于选择以小的数字开头。）

17.14 玩“4 位数乐透”。许多州有“4 位数乐透”，每天都会宣布一组中奖号码。中奖号码基本上相当于随机数字表里的一组 4 位数。如果你的 4 位数字和中奖号码的顺序完全相同，你就中奖了。奖金是由所有中奖的人平分。这样就有办法可以占些优势。

(a) 举例来说，中奖数字可能是 2 873 或是 9 999。说明为什么这两个结果出现的概率完全一样。（都是 $1/10\,000$ 。）

(b) 如果你问许多人，两个数字中哪一个比较有机会是随机选出的中奖数字，大部分人都会挑其中之一。用在本章学到的知识来判断，大部分人选的是其中的哪一个，并说明为什么。如果你选一个别人觉得比较不可能的数字，你赢的机会是一样大，可是你赢的钱比较多，因为选同样数字的人会比较少。

17.15 惊讶吗？你和学校分配给你的室友开始熟了。有一天在聊了很久之后，发现你们两人都有姐姐或妹妹名叫黛博拉。你应不应该很惊讶？说明你的答案。

17.16 沙奎尔的罚球。篮球运动员沙奎尔·奥尼尔(Shaquille O'Neal)在整个球季中，罚球差不多有一半会进。在今天的比赛当中，他的头三次罚球都进了。电视评论员说：“从今天的表现看来，沙奎尔的技巧有进步。”说明一下为什么这样判断沙奎尔罚球进步了是没有根据的。



17.17 长期结果。概率的解释，并不是要对不平衡的结果做补救，而是会把它淹没。假设掷铜板的头6次都得到反面，而之后掷的，都是一半正面，一半反面。（实际不大容易发生这样的完全平衡状况，但这个例子可以说明，为什么最开始的6个结果，会被之后的结果给淹没了。）掷了头6次之后，正面比例是多少？掷了100次之后，如果后面94次中有47个正面，则正面比例是多少？掷了1000次之后，如果后面的994次中有一半是正面，则正面比例是多少？掷了10 000次之后，如果后面的9 994次中有一半是正面，则正面的比例是多少？

17.18 “平均数定律”。棒球员汤尼·葛温(Tony Gwynn)在整个赛季中，安打比例大约是35%。在他连续6次挥棒都未击出安打之后，电视评论员说：“根据平均数定律，汤尼接下来一定会打出安打了。”这样说对吗？为什么？

17.19 冷冬要来了。有位气象学家预测下个冬天会比正常的冬天要冷，他说：“首先，过去的几个冬天都不怎么冷。虽然我们不应该用平均数定律，但是严冬应该要来了。”你觉得在讨论气候的时候，提出“根据平均数定律应该……”有没有道理？

17.20 脑袋有待开发的赌客。

- (a) 有一个赌客知道，在轮盘赌当中每转一次轮盘，出现黑色和红色的机会是均等的。他观察到红色连续出现5次之后，下一把就赌很多钱在黑色上面。被问到为什么时，他解释说：“因为根据平均数定律，应该要出现黑色了。”向这位赌客说明，这种推论方式有什么不对之处。
- (b) 听完你解释为什么在轮盘出现连5红之后，红色和黑色概率仍然相等以后，赌客改去玩一种扑克牌游戏。他连续被发给了5张红色的牌。他记得你说的话，假设下一张发到的牌，红色或黑色的概率相等。他的想法对还是错，为什么？

17.21 对风险的反应。参加中学足球队的比赛的死亡概率，在你参与的每一年当中，大约是0.1。你因为就读一所有石棉存在已10年的学校，因石棉而得癌症的概率差不多是0.000 005。如果我们不准学校用石棉，是不是也应该禁止高中足球比赛？简短说明你的立场。



17.22 对风险的反应。全国性的报纸诸如《今日美国》及《纽约时报》，刊登的坠机死亡的新闻，比撞车死亡的要多得多。在美国每年因车祸死亡的人数大约有 40 000。所有定期航班包括短航线在内，在近几年之内每年因坠机死亡的人数在 44—394 人之间。

(a) 媒体为什么给坠机事件较多关注？

(b) 媒体的报导如何有助于解释，为什么许多人以为飞行比开车危险？

17.23 概率没有这样说。掷铜板时的正面概率是 $1/2$ 。这个意思是说，当我们一直继续掷铜板的时候，正面的比例迟早会很靠近 0.5。但并不是说正面的次数会靠近所掷次数的一半。要知道为什么，想像一下我们掷铜板 100 次，1 000 次，10 000 次，掷出的正面比例都是 0.51。以上掷的每一回合，正面出现多少次？正面次数距掷出的总次数的一半差距有多少？

第 18 章

概率模型

正面还是反面，有人搞不清楚

玛莎：好，汤姆，这里有个新硬币。假如我掷它的话，你觉得正面概率会是多少？

汤姆：噢，我认为 60%。

玛莎：那你认为反面的概率呢？

汤姆：就 50% 好了。

玛莎：正面机会 60%，反面机会 50%，加起来的话正面或反面的概率有 110%。这样不合理。

汤姆：这个嘛，这些是我的个人概率，所以我要怎么定都可以。

玛莎：不对。正、反面概率一起考虑的时候，应该要合理才对。



玛莎说个人概率必须遵循合乎一般常识的规则才合理。而“合理”的概率不一定能描述我们掷许多次铜板会发生的状况。我们知道汤姆的个人概率，未必和铜板掷很多次后人头真正出现的比例一样。但就像我们会检查数据彼此之间是否相符一样，我们也坚持，概率彼此之间必须相符。同一个铜板的正面概率和反面概率，加起来不可以超过 1。事实上，加起来应该刚好是 1，除非我们认为有其他可能，比如铜板可能竖在那里。因为所有可能的结果一起考虑的话，概率一定恰好是 1。而这是我们在本章中要探讨的“概率规则”(rules of probability)之一。

概率模型

从 25—29 岁的女性当中随机选择一位，并记录她的婚姻状况。“随机”的意思是说，我们给了每位合乎条件的女性同样的人选机会。也就是说，我们选了一个大小为 1 的随机样本。任何一种婚姻状况的概率，就是所有 25—29 岁的女性当中，各种婚姻状况的人所占的比例，即如果我们抽取了许多位女性，也就会得到这个比例。以下就是这组概率：

婚姻状况	从未结婚	已婚	寡居	离婚
概率	0.386	0.555	0.004	0.055

上面这个表针对随机抽取一位年轻女性，并了解她的婚姻状况，提供了一个概率模型(probability model)。它告诉我们可能的结果有哪些(这里只有 4 种)，并给这些结果分配概率。这里的概率，就是每一种婚姻状况的女性所占比例。这样子应该很清楚，单身女性的概率，就是三类没有配偶女性的概率总和：

$$\begin{aligned}P(\text{单身}) &= P(\text{从未结婚}) + P(\text{寡居}) + P(\text{离婚}) \\ &= 0.386 + 0.004 + 0.055 = 0.445\end{aligned}$$



我们常常用 $P(\text{单身})$ 来当做“我们抽中的女性为单身的概率”的简写。你已见识到，我们的模型不是只对个别结果分配概率而已，我们还可以把个别结果的概率加起来，而求得任何一组结果的概率。

概率模型

一个随机现象的**概率模型**(Probability model)描述所有的可能结果，以及任意一组结果的概率要如何分配。我们有时把一组结果叫做一个**事件**(event)。

概率规则

因为上个例子里面的概率只是所有女性中各种婚姻状况的比例，所以会遵循比例的规则。以下是一些所有概率模型都该服从的基本规则：

- A. 任何概率都是介于 0 与 1 之间的数。所有的比例都是介于 0 与 1 之间的数，所以所有的概率也都是介于 0 与 1 之间的数。概率为 0 的事件永远不会发生，而概率为 1 的事件在每次试验时都会发生。概率为 0.5 的事件，长期下来有一半的时候会会发生。
- B. 所有可能的结果合并起来，概率应该是 1。因为每一次试验总会发生某个结果，所以所有可能结果的概率之和一定恰好是 1。
- C. 一个事件不发生的概率，等于 1 减去该事件发生的概率。如果某个事件发生的次数占有所有试验中的 70%，则它在另外的 30% 就没有发生。一个事件发生的概率，及该事件不发生的概率，加起来必定是 100%，也就是 1。
- D. 如果两个事件当中没有共同的结果，则该两个事件中至少有一个会发生的概率，是该两事件个别概率的和。如果一个事件发生于所有试验中的 40%，另一个事件发生于所有试验中的 25%，而该两个事件不可能同时发生，则至少其中一个事件会发生的次数会占有所有试验的 65%，因为 $40\% + 25\% = 65\%$ 。



例1 年轻女性的婚姻状况

再来看看年轻的女性在各种婚姻状况的概率。4个概率当中的每一个都是介于0与1之间的数。加起来是

$$0.386 + 0.555 + 0.004 + 0.005 = 1$$

这个概率分配满足规则A及B。对每个个别结果分配概率的方法，只要满足规则A和B的，就是“合法的”。这是说，这样的一组概率是有意义的。此时规则C和D会自然成立。以下是应用规则C的例子：

根据规则C，我们抽到的女性是为单身的概率为

$$P(\text{单身}) = 1 - P(\text{已婚}) = 1 - 0.555 = 0.445$$

这是在说，如果其中的55.5%已婚，则剩下的44.5%就是单身。规则D说，你也可以把3种不同的单身状况的概率相加，求出女性为单身的概率。我们之前就是这样算的，算出来的结果和上面的一样。

例2 掷骰子

掷骰子是在赌场里输钱的一个很普遍的方法。当我们掷两颗骰子，并依序(第一颗骰子、第二颗骰子)记录朝上那面的点数时，总共会得到36种可能结果。图18.1中展示了这些结果。我们应该怎样分配概率呢？

赌场的骰子是很谨慎的制作出来的。为使每一面都一样重，有点的地方并不是凹陷的，而是用白色塑胶填平的，而且白色塑胶的密度和制作骰子本身的红色塑胶密度相同。对于赌场的骰子来说，对图18.1中36个结果都分配一样的概率是合理的。因为这36个概率的和必须是



1(规则 B), 每个结果的概率必定是 $1/36$ 。

我们感兴趣的是骰子朝上的面的点数之和。这个和是 5 的概率是多少?“掷出 5 点”这个事件包含 4 个结果, 而它的概率是这 4 个结果的概率之和:

$$\begin{aligned}
 P(\text{掷出 5 点}) &= P\left(\begin{array}{|c|c|} \hline \cdot & \cdot\cdot\cdot \\ \hline \end{array}\right) + P\left(\begin{array}{|c|c|} \hline \cdot\cdot & \cdot\cdot \\ \hline \end{array}\right) + P\left(\begin{array}{|c|c|} \hline \cdot\cdot\cdot & \cdot \\ \hline \end{array}\right) + P\left(\begin{array}{|c|c|} \hline \cdot\cdot\cdot & \cdot \\ \hline \end{array}\right) \\
 &= \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} \\
 &= \frac{4}{36} = 0.111
 \end{aligned}$$

概率规则只告诉我们哪些概率模型有意义, 并没说概率的分配是否正确, 是不是真的能描述长期状况。例 2 中的概率对赌场骰子来说是正确的。点的部分挖空的便宜骰子各面并不平衡, 因此例 2 的概率模型就不能描述这类骰子的状况。

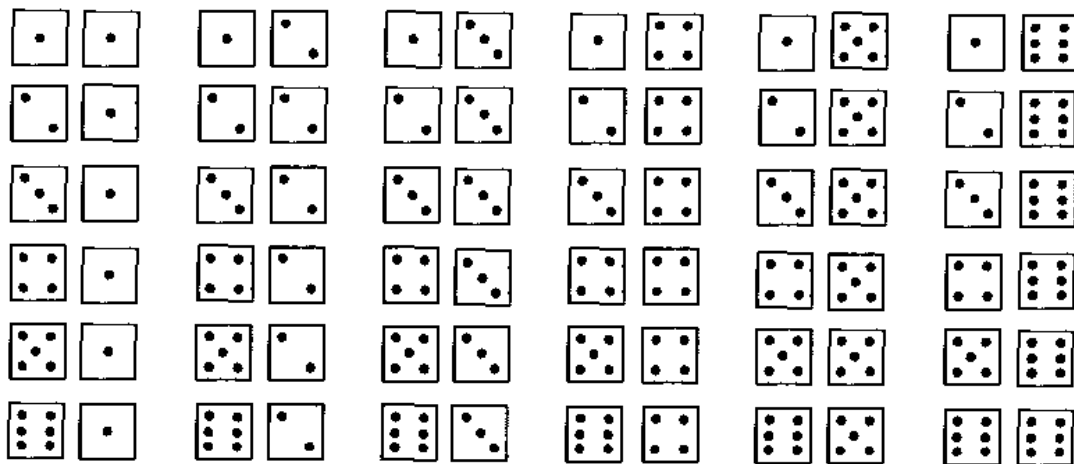


图 18.1 掷两个骰子的 36 种可能结果

至于个人概率又如何呢?如同汤姆所说的:“这个嘛, 这些是我的个人概率, 所以我要怎么定都可以。”我们不能说不符合规则 A 和 B 的个人概率一定不对, 但是我们可以说它们是不协调的。也就是说, 没有合理的方式可以把它们放在一起同时考虑。所以我们通常坚持, 对某一随机现象的所有结果所分配的个人概率, 必须符合规则 A 和 B。也就是说同一组规则对两种概率都适用。



抽样的概率模型

从总体选取随机样本，并计算像样本比例这样的统计量，当然是随机现象。统计量的分布告诉我们，它的可能值有哪些，以及每个值出现的频率。这听起来非常像是概率模型。

例 3 抽样分布

抽取一个 1 523 位成人的简单随机样本(SRS)。问其中一个人在过去 12 个月当中有没有买过乐透彩。答“有”的比例如下：

$$\hat{p} = \frac{\text{答“有”的人数}}{1\,523}$$

这就是我们的样本比例 \hat{p} 。重复这个步骤 1 000 次，并从所得的 1 000 个样本，算出 1 000 个样本比例 \hat{p} 。图 18.2 的直方图所展示的，是当总体中买乐透彩的真正比例为 60% 时，1 000 个样本比例的分布。随机样抽样的结果当然是随机的：我们没法子预知一个样本的结果，但是从图可以看出来，许多样本的结果摆在一起，是有规则形态的。

我们曾在第 13 章里面见过图 18.2。事实上早在第 13 章和第 11 章中，我们就见过这个图里的直方图部分了。这样不断出现也提醒了我们，重复抽取随机样本所出现的规则形态，是统计的重要概念之一。图里的正态曲线，是直方图一个合理的近似。直方图是这里特定的 1 000 个 SRS 的结果。你就把正态曲线想成是若我们永不间断的从这个总体抽取 SRS 时，所会得到的理想化形态。这和概率概念完全一样，概率就是长久下来我们会见到的模式。正态曲线对随机抽样的结果分配概率。

这个正态曲线的平均数为 0.6，标准差大约 0.012 5。68-95-99.7 规则的“95”部分说，所有样本当中有 95%，其 \hat{p} 会落在平均数左右 2 个标准差范围之内。也就是在 0.6 ± 0.025 的范围内，即介于 0.575 与 0.625 之间。对这个事实我们已经有更确实的语言可以表达：一个样本中有 57.5%~62.5% 的人会答“有”的概率是 0.95。“概率”这个词是说，我们谈的是长期下来，有了许多样本时会发生什么事。



从一个很大的样本算出来的统计量，会有非常多的可能值。对每一个可能结果分配概率这档子事，对于4种婚姻状况或者掷一对骰子的36种结果来说都没问题，但是在可能结果有几千种的情况下就不太方便了。例3用了不同的方法：利用正态密度曲线底下的面积，来对区间形式的结果分配概率。密度曲线底下的面积是1，这点和总概率为1刚好配合。图18.2中正态曲线底下的总面积为1，而在0.575和0.625之间的面积是0.95，这就是一个样本所得结果会落入该区间的概率。当利用正态曲线计算概率时，你可以用68-95-99.7规则来算，也可以利用表B中的正态曲线百分位数。这些概率都符合规则A到D。

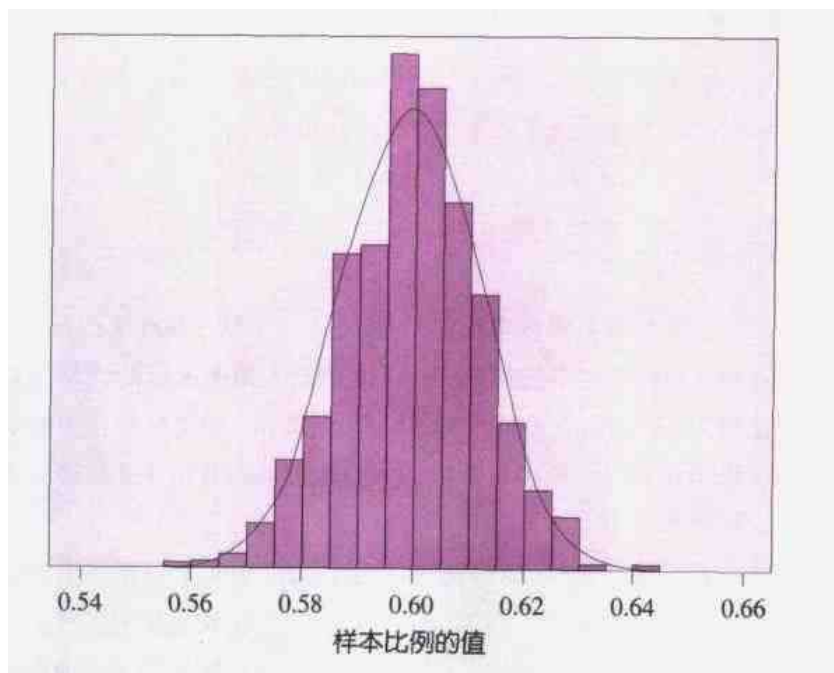


图 18.2 在一个有 60% 的人会给肯定答复的总体中，抽取大小为 1 523 的 SRS，所得样本比例 \hat{p} 的抽样分布。直方图呈现的是 1 000 个样本的分布。正态曲线是描述很多很多个样本所得结果的理想化形态

• 抽样分布

一个统计量的抽样分布(sampling distribution)告诉我们，从同一总体重复抽样时，统计量会有些什么样的值，以及每个值出现的频率。

我们把抽样分布看成是对统计量的可能值分配概率。因为通常可能值有许多，所以抽样分布常常是用诸如正态曲线的密度曲线来描述。

**例 4 你赞成赌博吗?**

某意见调查问 501 位十几岁年轻人的 SRS 问题：“一般来说，你赞成还是反对合法的赌博或下注？”假设事实上在十几岁的人当中，被问到时会有恰好 50% 的人答“赞成”。（这个数字和通过数种调查所显示的真正比例很接近。）该调查的统计学家告诉我们，在不同的样本当中，答“赞成”的样本比例会一直变，其分布遵循平均数 0.5、标准差 0.022 的正态分布，这就是样本比例 \hat{p} 的抽样分布。

根据 68-95-99.7 规则，该调查会得到一个其中少于 47.8% 的人说“赞成”的样本的概率是 0.16。图 18.3 显示出怎样可以从抽样分布的正态曲线得到这个结果。

例 5 利用正态分布百分位数*

例 4 当中的抽样调查抽到一个有 52% 以上的人说“赞成”的样本，概率为多少？因为 0.52 并不是与平均数相差 1、2 或 3 个标准差，所以没法子利用 68-95-99.7 规则。我们会有表 B 的正态分布百分位数。

要利用表 B，先得把结果 $\hat{p} = 0.52$ 减掉分布的平均数，再除以标准差，来转换成标准计分：

$$\frac{0.52 - 0.5}{0.022} = 0.9$$

现在来看表 B。标准计分 0.9 是正态分布的第 81.59 百分位数。这个意思是说，调查结果得到较小的比例的概率是 0.8159。根据规则 C（或者就用曲线底下总面积为 1 这个事实），会有 52% 或更多人赞成的概率就是 0.1841。图 18.4 把这个概率用正态曲线底下的面积表示。

* 例 5 为选读

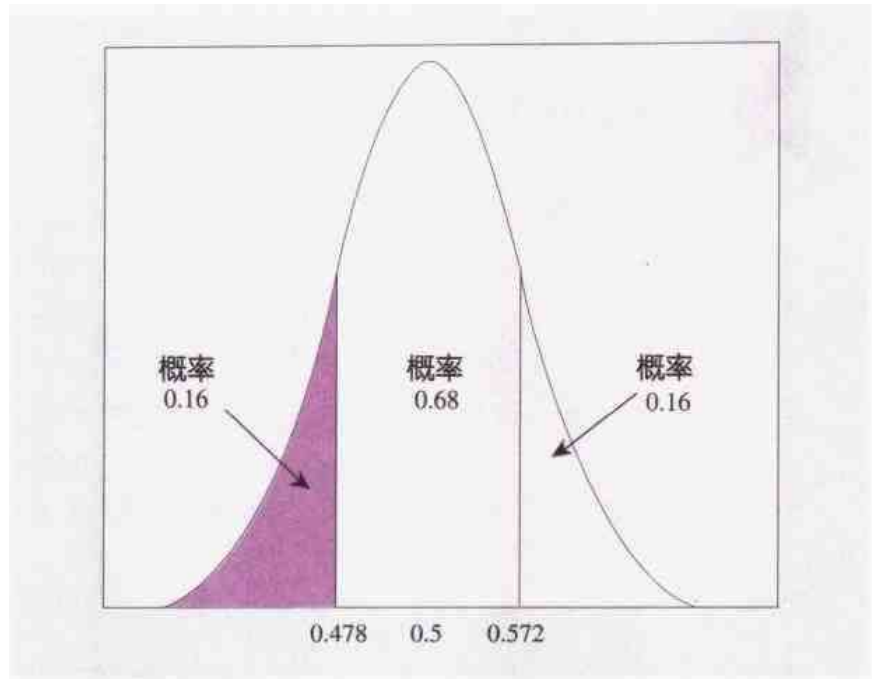


图 18.3 例 4 的正态抽样分布。因为 0.478 是在平均数之下 1 个标准差的位置，所以曲线之下在 0.478 以左的面积是 0.16

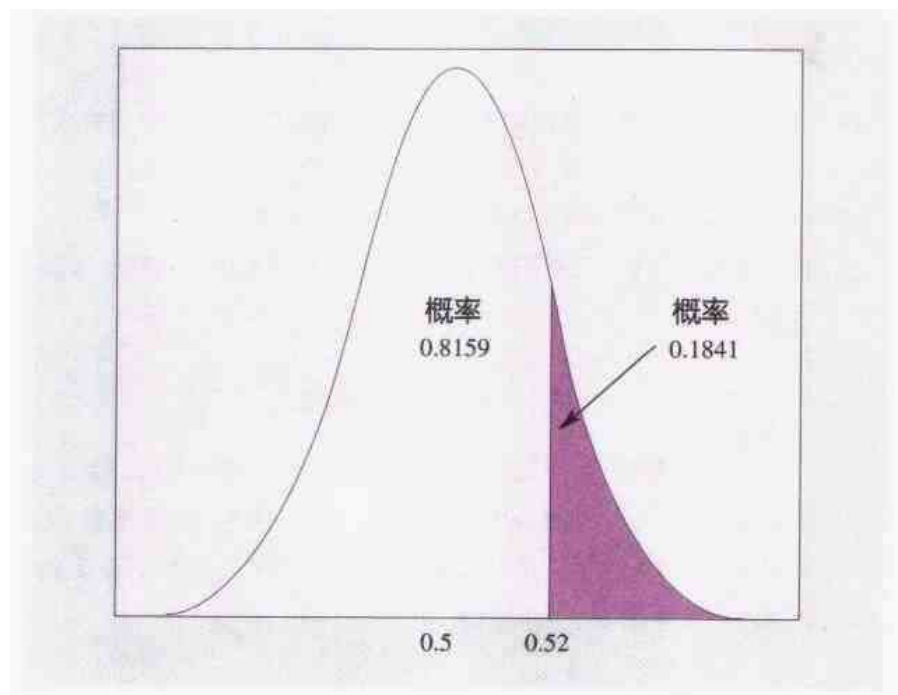


图 18.4 例 5 的正态抽样分布。0.52 这个结果的标准计分是 0.9，所以根据表 B，曲线之下 0.52 以左的面积是 0.8159



本章重点摘要

我们用**概率模型**来描述随机现象，方法是说明有哪些可能结果，以及要怎样分配概率给这些结果。有两种简单方式可呈现出概率模型。第一种是分配概率给每一个个别结果，这些概率必须是介于 0 与 1 之间的数(规则 A)，而且加起来要恰好是 1(规则 B)。若要找某个事件的概率，只要把组成该事件的结果的概率加起来即可。

第二种概率模型是以某一**密度曲线**之下的面积来分配概率，比如像正态曲线。总概率是 1，因为曲线底下的总面积是 1。这一类的概率模型通常用来描述统计量的抽样分布。这是指从同一总体抽许多样本所得到统计量的值形成的形态。

所有“合法的”概率分配，不论是根据数据所得还是个人概率，都遵循同样的**概率规则**。因此概率的计算方法都是一样的。



第18章 习题

18.1 力争上游。一位研究丹麦人民社会地位变动状况的社会学家发现,若父亲的社会地位不高,则儿子日后的社会地位仍不高的概率是 0.46。则儿子会上升到较高地位的概率是多少?

18.2 死因。在美国的政府资料中,对每一个在美国发生的死亡事件,都会记录一个单一死因。资料显示,随机选取一个死亡事件,死于心血管疾病(主要是心脏)的概率为 0.45,于癌症的概率为 0.23。死因为心血管疾病或癌病的概率是多少?死因既非心血管疾病亦非癌症的概率是多少?

18.3 加拿大的土地。在加拿大随机选取 1 英亩的地。这块地是森林的概率为 0.35、是牧场的概率为 0.03。

(a) 选中的地不是森林的概率是多少?

(b) 是森林或牧场其一的概率是多少?

(c) 在加拿大随机选取的 1 英亩地既不是森林又不是牧场的概率是多少?

18.4 老公有没有做家务?一项民意调查访问了 1 025 位女性的随机样本。样本中的已婚女性被问到她们的老公有没有公平分摊家务。结果如下:

结果	概率
做得比他分内的还多	0.12
做了他分内的家务	0.61
没有做到他分内该做的	?

这些比例是随机选一位已婚女性,并问她意见的随机现象的概率。

(a) 选中的女性会说她老公没有做到分内该做的家务的概率,必定是多少?为什么?

(b) “我觉得我老公至少做了他分内该做的”这个事件包含了头两个结果。这个事件的概率是多少?









结果	概率			
	模型 1	模型 2	模型 3	模型 4
	1/7	1/3	1/3	1
	1/7	1/6	1/6	1
	1/7	1/6	1/6	2
	1/7	0	1/6	1
	1/7	1/6	1/6	1
	1/7	1/6	1/6	2

图 18.5 掷一个骰子的 4 种概率模型, 对照习题 18.5

18.5 掷骰子。图 18.5 中呈现了好几种对一个骰子的正面的概率分配, 对一个特定骰子来说, 哪一种概率分配是正确的, 必须靠掷这颗骰子很多次才会知道。然而这儿的几种概率分配当中有的并不合法, 也就是说并不符合概率规则。哪些合法哪些不合法? 对于不合法的模型要说明是哪里不对。

18.6 高中成绩排名。从大一学生当中任选一人, 询问高中成绩的排名高低。以下是根据对大一学生做的大规模抽样调查的比例所得的概率:

排名	最高 20%	次高 20%	最中间 20%	次低 20%	最低 20%
概率	0.41	0.23	0.29	0.06	0.01

- (a) 这些概率的和是多少? 为什么你会预期和会是这个值?
- (b) 一个随机选取的大一学生, 在高中时成绩没有在班上前 20% 的概率是多少?
- (c) 大一学生在高中时的成绩在班上前 40% 的概率是多少?

18.7 四面体骰子。心理学家有时会利用四面体骰子来研究我们对机遇事件的直觉。一个四面体(见图 18.6)是有四个面的金字塔, 每一面都是等边三角形。把一个四面体骰子的四个面分别点上 1、2、3 及 4 个点。掷一个这样的骰子并记录朝下那一面的点数, 对掷出的结果提出一个概率模型。说明为什么你认为你的模型至少很接近正确的状况。

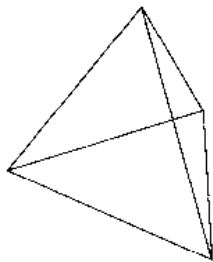


图 18.6 四面体。习题 18.7 和 18.9 都是讨论这种形状的骰子

18.8 出生顺序。 对夫妇打算生 3 个小孩。小孩性别以及出生顺序总共可以排出 8 种可能。举例来说，女女男代表头两个孩子是女孩，而第三个是男孩，全部 8 种可能的概率都几乎一样(我们把几乎一样视为相同)。

(a) 把 3 个孩子性别的可能安排都写出来。其中每一种安排的概率是多少?

(b) 该对夫妇的 3 个孩子包括两女一男的概率是多少?

18.9 又见四面体骰子。 习题 18.7 里面已经说明过四面体骰子是什么东西。写出掷两个四面骰的概率模型。也就是说，写出所有可能结果并给每个结果分配概率。(可参考例 2 及图 18.1)。两个骰子朝下的面点数和为 5 的概率是多少?

18.10 轮盘。 一个轮盘分成 38 格，编号为 0、00 以及 1—36。0 和 00 这两格是绿色，其他 36 格中有 18 格是红色，18 格是黑色。庄家转动转盘，同时反方向把一个小球沿盘缘滚上转盘。转盘的水平经过仔细校正，使得当转盘慢下来时，球落在任一格的概率相同。赌客可以下赌注于多种数字和颜色之组合。

(a) 38 种可能结果中的任一种的概率是多少?说明你的答案。

(b) 如果你赌“红色”，则当球落在红色格子里时你就赢了。赢的概率是多少?

(c) 格子的号码都画在一个桌面上，赌客就把筹码摆上去。桌面上有一行的数字都是 3 的倍数，也就是 3、6、9……36。你以格形压注(column bet)，只要你压注的这一行当中，有任一个号码中了，你就赢了。你赢这一把的概率是多少?

18.11 M&M 巧克力的颜色。 如果你从一袋 M&M 里面随意抓一颗



糖出来，它的颜色有6种可能。抓到任一特定颜色的概率，和全部糖果中每一种颜色占的比例有关。

(a) 以下是任选一颗纯 M&M，每种颜色的概率：

颜色	褐	红	黄	绿	橘	蓝
概率	0.3	0.2	0.2	0.1	0.1	?

抽到蓝色糖果的概率必定是多少？

(b) 花生 M&M 的概率有点儿不一样。像下面这样：

颜色	褐	红	黄	绿	橘	蓝
概率	0.2	0.1	0.2	0.1	0.1	?

任选一颗花生 M&M，会选到蓝色的概率是多少？

(c) 一颗纯 M&M 是红、黄、橘其中一色的概率是多少？一颗花生 M&M 是这三色之一的概率又是多少？

18.12 合法概率？对以下3种状况中对个别结果的概率分配，决定何者合法(满足概率规则)。对于不合法者，请写出明确的理由。

(a) 掷一个铜板， $P(\text{正面}) = 0.55$ ， $P(\text{反面}) = 0.45$ 。

(b) 掷两个铜板， $P(\text{正正}) = 0.4$ ， $P(\text{正反}) = 0.4$ ，

$P(\text{反正}) = 0.4$ 及 $P(\text{反反}) = 0.4$ 。

(c) 纯 M&M 并不是一直有像习题 18.11 中的颜色组合。以前没有红色也没有蓝色。黄褐色的概率是 0.1，其他 4 种颜色的概率和习题 18.11 中的一样。

18.13 女性意见调查。假设所有的成年女性中，有 47% 认为她们没有足够可以运用的个人时间。一项抽样调查访问了 1 025 位随机选择的女性，并记录了自己觉得没有足够个人时间的样本比例。如果不断重复做这项调查，样本比例这个统计量的值会随着样本而变。其抽样分布近似正态，平均数为 0.47，标准差大约为 0.016。画出这条正态曲线，并用它来回答下面问题：

(a) 总体的真正比例是 0.47。所有样本结果的中间 95% 会落在什么范围？



- (b) 调查时所抽到的样本，其中少于 45.4% 的人说没有足够个人时间的概率多大？

18.14 我们会怎么死？假设所有成人当中有 70% 认为因枪击死亡的事件以后会增加。一项意见调查计划抽取 1 009 位成人的 SRS，询问他们对枪支暴力的意见。如果我们从这同一个总体抽取许多个总体，认为因枪击而死亡的事件会增加的人所占比例，会跟着样本而变。样本比例的抽样分布近似于平均数为 0.70，标准差约为 0.014 的正态分布。画出这条正态曲线，并用它来回答下面问题：

- (a) 抽到的样本当中，超过 72.8% 的人认为枪击致死事件会增加的概率是多少？
(b) 抽到的样本之样本比例会和真正比例(70%)相差至少 2.8% 的概率是多少？

18.15 女性意见调查(此题可略过)。在习题 18.13 的架构下，抽到的样本中有超过 51% 的女性觉得个人时间不够的概率是多少？(要用表 B。)

18.16 我们会怎么死？(此题可略过)？在习题 18.14 的框架下，抽到的样本中有不到 66.5% 的人认为以后枪击致死事件会增加的概率是多少。(要用表 B。)

18.17 你慢跑吗？一项意见调查问 1 500 位成人的 SRS：“你慢跑吗？”假设(这点大致正确)总体中的慢跑比例是 $p = 0.15$ 。在众多样本当中，回答“是”的样本比例 \hat{p} 会大致是平均数 0.15，标准差 0.009 的正态分布。画出这条正态曲线，并用它来回答下列问题：

- (a) 为数众多的样本当中有多少百分比，其样本中之慢跑比例会小于或等于 0.15？仔细解释为什么这个百分比就是 \hat{p} 会小于或等于 0.15 的概率。
(b) \hat{p} 的值在 0.141—0.159 之间的概率是多少？(用 68-95-99.7 规则。)
(c) 现在用概率规则 C: \hat{p} 不在 0.141—0.159 之间的概率是多少？

18.18 申请大学入学。你问了一个含 1 500 位大学生的 SRS，他们有没有申请别的学校。假设事实上大学生中有 35% 曾申请其他学校



(这很接近事实)。样本中会说“有”的比例 \hat{p} 其抽样分布大致是平均数 0.35、标准差 0.01 的正态分布。画出这条正态曲线，并用它来回答下列问题：

- (a) 用一般用语来说明，抽样分布对于我们的样本结果提供了什么信息？
- (b) 众多样本当中，有多少百分比会有大于 0.37 的 \hat{p} 值？(用 68 - 95 - 99.7 规则。)用一般用语解释，为什么这个百分比就是样本结果大于 0.37 的概率。
- (c) 你的样本的 \hat{p} 值会小于 0.33 的概率是多少？
- (d) 用概率规则 D：你的样本结果不是小于 0.33 就是大于 0.35 的概率是多少？

18.19 建构抽样分布。让我们从很小的一个总体，抽一个极小的样本，来说明抽样分布的概念。总体是 10 个学生在一项考试中的分数：

学生	0	1	2	3	4	5	6	7	8	9
分数	82	62	80	58	72	73	65	66	74	62

我们感兴趣的参数是这个总体的平均分数。样本是抽自总体，大小为 $n=4$ 的 SRS。因为学生已经用 0 到 9 当代码，所以从表 A 中抽一个随机数字，就相当于样本中的一个学生。

- (a) 算出总体中 10 个分数的平均。这是总体平均数。
- (b) 利用表 A 从这个总体抽出一个大小为 4 的 SRS。把这 4 个学生的分数常做你的样本，并计算平均数 \bar{x} 。这个统计量是总体平均数的一个估计。
- (c) 用表 A 的不同部分重复比过程 10 次。画出这 10 个 \bar{x} 值的直方图。你正在建造 \bar{x} 的抽样分布。你的直方图其中心和在(a)中所算出来的总体平均数接近吗？

第 19 章

模拟

提高收费站的通行速度

我们对于在桥梁、隧道或高速公路收费站的长龙都很熟悉并且痛恨。统计没有办法减少要通过的车辆数目，但是可以帮大家加快通过的速度。

假设我们现在必须为一条新隧道设计收费站。两个方向的车流都要在同一个地点付费，驾驶员可以选择付现金、信用卡或者一种不必停车的电子付费系统。我们会需要几个收费车道呢？应该要让每个车道都接受所有付费方式，还是专攻一种？旅游旺季时，多数游客会付现金，使用当地电子付费系统的驾驶员变少，这时会发生什么状况？当某个收费车道的车辆增加一倍时，等候时间常常变成四倍或更糟，



此时要考虑高峰时段的车辆堵塞，会不会阻碍了主要道路的交通？

这整个情况过于复杂，不是光想就能解决的，无法以任何一组我们可用来找出的解答的数学方程式来描述这个状况。我们能做的是请出我们的随机数字产生器，然后试着模仿驾驶员的行为，这种模仿的正式名称是模拟(simulation)。汽车和卡车开到收费站的时间是随机的，但是我们可从过去的资料看出到达时间和车辆种类的概率分布。驾驶员会选择不同的付费方式，这又可以用更多的概率分布来描述(对当地人的事、游客的车及卡车，分布都会不同)。不同的驾驶员急着要通过的程度又不同，这又是另一个概率分布。收费站服务一位驾驶员所花的时间也是随机的，遇到手上拿着 50 美元大钞又要收据的老兄，就会花比较多的时间。

现在我们可以电脑屏幕上观察收费站的运作情形。车流到达时，等候的时间和车队的长度会随着增减。旅游旺季时，高峰时间的队伍是不是经常都太长了？加一个收费车道会怎样？如果增加的这个车道只收现金又会怎样？假如我们把进站距离拉长，让驾驶员有较多时间找出最短的队伍，情况会如何？如果，万一交通量增加一倍，那会变成什么状况？当然不可能发生什么好事，但也许我们可以有办法不让现金车道的队伍太长，因而挡住要去排电子收费车道的车辆。

模拟像我们的收费站这样，内部有复杂相关性的系统，已经有一种标准做法。细节多得可怕，但是观念却很简单。模拟背后的观念，就是本章的主题。

概率从何而来？

掷铜板时得到正面和反面的概率都很接近 0.5。原则上来说，这些概率是根据掷很多次的数据得来的。乔对于明年谁会赢足球超级杯的个人概率，却是根据乔的个人判断得来的。掷一个铜板 10 次，而会出现连续 3 次正面的概率又要怎么找呢？我们可以利用描述掷铜板



真正随机的数字

对于要求“纯正”随机数字的人，兰德公司 (RAND Corporation) 很久以前就出版了《一百万个随机数字》(One Million Random Digits)。书中列出了 1 000 000 个数字，它们是由很复杂的物理随机系统产生，是真正的随机数字。有一位兰德公司的工作人员曾告诉我，这还不是兰德公司出版的书当中最无聊的一本……

情况的模型，计算出这个概率。也就是说，一旦我们根据数据，找到掷铜板的概率模型后，就不需要在每次要找一个新事件的概率时，都再重新开始。

用概率模型的一大优点，是让我们可以只先对一些如“掷一次铜板得到正面”的简单事件分配概率，就能够计算一些复杂事件的概率，不论概率模型反映的是根据数据得到的概率还是个人概率，上面的优点都成立。

但不幸的是，计算概率所需要用到的数学常常很难。还好有科技拯救我们；只要我们有概率模型，就可以用电脑来模拟重复许多次的状况。这样做比算数学容易，更比在真实世界中执行许多次的重复要快得多。你可以把利用模拟求概率的方法，和利用电脑控制的飞行模拟器来练习飞行做个比较。两种模拟都有很多人用，也都有共同的弱点：模拟的效果取决于你给的模型，若模型不恰当，模拟结果就不可能好。飞行模拟器使用的是飞机会如何反应的软件“模型”；模拟概率时使用的是概率模型。我们会用我们的旧识，表 A 中的随机数字，来启动我们的模型。

• 模拟

利用随机数字表或者电脑软件中的随机数字，来模仿机遇现象，就叫做模拟(simulation)。

我们讨论模拟，一部分是因为设计收费站的工程师的确是用这个办法来算概率的，一部分也因为模拟会迫使我们仔细思考概率模型的意义。我们会做困难的部分，也就是建立模型；而容易的部分，例如叫电脑去重复 10 000 次，就留给真正需要用到最后算出来的概率的人。

模拟入门

一旦有了可靠的概率模型，模拟是找出复杂事件发生概率的有效工具，我们可以利用随机数字，很快就模拟出多次重复的结果。一个事件在这些重复结果中发生的比例，迟早会接近它的概率，所以模拟



可以对概率做适当的估计。要学习模拟的艺术，最好的方法就是连续看几个例子。

例1 如何执行模拟

掷一个铜板 10 次。结果中会出现至少 3 个连续正面或是至少 3 个连续反面的概率是多少？

第1步：提出概率模型。我们的掷铜板模型含有两部分：

- 每一次掷，正面和反面的概率各为 0.5。
- 投掷之间，彼此是独立的。也就是说，知道某一次掷出的结果，不会改变任何其他次所掷结果的概率。

第2步：分配随机数字以代表不同的结果。表 A 随机数字表中的数字会以符合第 1 步之概率的方式，来代表各种结果。我们知道：表 A 中的每一个数字会是 0、1、2、3、4、5、6、7、8、9 中任一个的，概率都是 0.1，而且表中数字之间是互相独立的。以下是针对掷铜板结果分配数字的方法之一：

- 每个数字模拟掷一次铜板的结果。
- 奇数代表正面，偶数代表反面。

这样子的分配可行，是因为 5 个奇数使正面概率恰好是 5/10。表中的连续数字可模拟多次独立的投掷。

第3步：模拟多次重复。10 个数字模拟 10 次投掷，所以表 A 中的 10 个连续数字模拟了一组掷铜板 10 次的状况。在表 A 中读取许多组的连续 10 个数字，就模拟了重复多次的状况。别忘了，在每次重复时，要记录我们开心的事件（至少连续 3 个正面或反面）有没有发生。以下是头 3 组重复的结果，是从表 A 中的列 101 开始的。我把所有连续 3 个或更多个的正面或反面底下都画了线。

	第一组										第二组									
数字	1	9	2	2	3	9	5	0	3	4	0	5	7	5	6	2	8	7	1	3
正/反面	正	正	反	反	正	正	正	反	正	反	反	正	正	正	反	反	反	正	正	正
连续 3 个					有										有					



第三组										
数字	9	6	4	0	9	1	2	5	3	1
正/反面	正	反	反	反	正	正	反	正	正	正
连续 3 个	有									

在表 A 中继续下去，我一共做了 25 组；其中 23 组有连续至少 3 个或更多的正面或反面。所以我们用比例来估计这个事件的概率为

$$\text{估计概率} = \frac{23}{25} = 0.92$$

当然，光做 25 次，不足以让我们对估计值的正确度有信心，可是现在既已了解模拟是怎么做的，我们可以叫电脑重复做好几千次。经过长串的模拟(或者确实的数学计算)，得到真正的概率大约是 0.826。大部分的人认为连续正面或反面不太容易发生，但连我这个很小规模的模拟结果所显示的，掷 10 次铜板大部分时候都会出现连续 3 个同样的结果，这已足以修正我们直觉的错误。

一旦你对模拟有一些经验以后，会发现整个过程最困难的部分，通常就是建立概率模型(第 1 步)。虽然掷铜板这个例子可能不大吸引你，例 1 中的模型却能代表许多概率问题，因为它是由许多独立的试验(掷铜板)构成，而每次试验都有一样的可能结果及概率，投篮 10 次和观察 10 个孩子的性别，也有类似的模型，也可用几乎一样的方法模拟。这个模型新的部分就是：各次试验彼此都是独立的，这点假设可以简化我们的模拟工作，因为可以用完全一样的方法，模拟掷 10 次铜板当中的每一次结果。

• 独立

如果“知道两个随机现象其中之一的结果”，并不会改变另一个结果的概率，就称这个两随机现象是独立的。

独立就和概率的其他性质一样，一定要重复观察很多次，才能证实。重复投掷铜板应该是独立的(铜板没有记忆)，经过观察后也发觉



事实确是如此。但要说一个篮球员的前后投球之间彼此独立，就不那么可信，不过观察显示，它至少很接近独立。

第2步(分配随机数字)根据的是随机数字表的性质。以下是这个步骤的一些例子。

例2 分配数字以执行模拟

(a) 从就业人口占70%的一群人当中，随机选一个人。一个数字代表一个人：

0, 1, 2, 3, 4, 5, 6 = 已就业者

7, 8, 9 = 未就业者

(b) 从就业人口占73%的一群人当中，随机选一个人。现在得用两个数字来模拟一个人：

00, 01, 02...72 = 已就业者

73, 74, 75...99 = 未就业者

我们把100个二位数的数对中的73个分配给“就业者”，以便得到0.73的女性。如果用01、02...73来代表“就业者”，也同样正确。

(c) 从50%有就业，20%失业以及30%不属于劳动人口的一群人中随机选一人。现在有3种可能的结果了，但是原则是一样的。一个数字模拟一个人：

0, 1, 2, 3, 4 = 就业者

5, 6 = 失业者

7, 8, 9 = 不属于劳动人口

更复杂的模拟

随机模型的建造和模拟是现代科学强有力的工具，并且不需要高深的数学就可以了解。还不只是这样，只要你自己试试模拟随机现象



他是技术好
还是运气好?

当一位棒球员有 300 的打击率时，大家都喝采。打击率 300 的打击者，在所有打数中，有 30% 击出安打。整年打击率为 300 会不会只是运气？一般大联盟的球员，一个球季大约有 500 次打数，打击率大约 260。打者的各个打数之间似乎是独立的。根据这个模型，我们可以计算或者模拟出打击率为 300 的概率。这个概率约为 0.025。在 100 位一般的大联盟打击者中，一年会有两、三个因为运气好而打出 300 的打击率。

几次，就会加强你对概率的了解，比读很多页我的文章还有用。现在我们心里有两个目标：一是要了解模拟本身，一是为更了解概率而来了解模拟。我们来看看两个较复杂的例子。第一个例子的试验，彼此仍然是独立的，然而不像我们掷铜板 10 次那样，有固定的试验次数。

例 3 我们要女儿

一对夫妇计划生孩子生到有女儿才停，或生了 3 个就停止。他们会拥有女儿的概率是多少？

第 1 步：概率模型和掷铜板的相似：

- 每一个孩子是女孩的概率是 0.49，是男孩的概率是 0.51。（没错，新生婴儿中男孩比女孩多。但男孩在婴儿时期的死亡率较高，所以两性的人数差别在发生影响之前就差不多扯平了）
- 各个孩子的性别是互相独立的。

第 2 步：分配数字也很容易。用两个数字模拟一个孩子的性别。我们把 100 个数对中的 49 个分配给“女孩”，另外的 51 个分配给“男孩”；

00, 01, 02, ..., 48 = 女孩

49, 50, 51, ..., 99 = 男孩

第 3 步：要模拟这个生孩子策略的 1 个回合，我们要从表 A 当中读取一对一对的数字，直到这对夫妇有了女儿，或已有 3 个孩子。模拟一回合所需要的数对数目，要看这对夫妇多快生出个女儿而定。以下就是我们的 10 次模拟，用的是表 A 的列 130。为了解释随机数字，我在它们底下写女代表女孩，男代表男孩，用空格把每次的模拟隔开，并在每次模拟的下方，写“+”号代表有生女孩，“-”号代表没有。



6905	16	48	17	8717	40	9517	845340	648987	20
男女	女	女	女	男女	女	男女	男男女	男男男	女
+	+	+	+	+	+	+	+	-	+

在这 10 次重复中，有 9 次生了女孩。我们对于用这个策略会得到女孩的概率的估计是：

$$\text{估计概率} = \frac{9}{10} = 0.9$$

用数学可以计算出来，如果我们的概率模型正确的话，会有女孩的真正概率是 0.867。

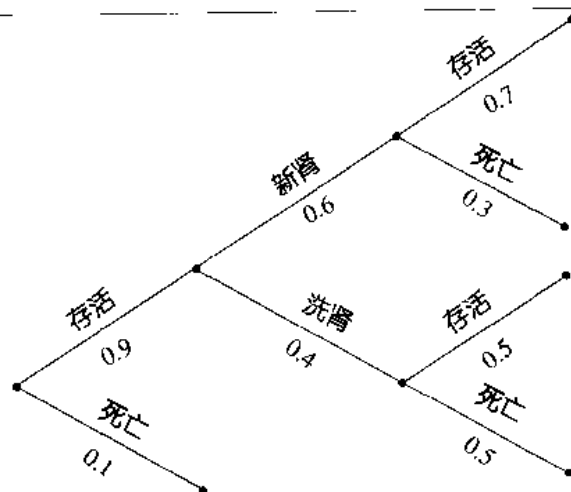
我们的模拟答案相当接近了。除非这对夫妇运气很不好，他们应该可以成功拥有一个女儿。

我们的最后一个例子当中有分阶段，阶段之间彼此不独立。也就是说，一个阶段的概率，和前一阶段的结果有关。

例 4 肾脏移植

莫理斯的肾脏不行了，他在等待肾脏移植。他的医师提供了和他情况类似的病人资料，撑过移植手术的占 90%，另外 10% 会死亡。在手术后存活的人中有 60% 移植成功，另外的 40% 还是得回去洗肾。五年存活率对于有新肾的人来说是 70%，对于回去洗肾的人来说是 50%。莫理斯希望知道，他能活过五年的概率。

第 1 步：图 19.1 中的树图(tree diagram)把这些信息组织了起来，用图的形式来表达出概率模型。树图显示出 3 个阶段，以及每阶段的可能结果及概率。树的每一条路径的终点，不是存活超过五年就是在五年内已死亡。要模拟出莫里斯的命运，我们必须模拟 3 阶段中的每一个阶段。第 3 阶段的概率，和第 2 阶段的结果有关。



阶段1 阶段2 阶段3
移植 移植成功? 存活五年?

图 19.1 例 4 概率模型的树图。每一个分枝点就是一个新阶段的开始，其结果和概率都写在树枝上。此模型的每一个模拟阶段，是从分枝点走到某一个端点

第 2 步：以下是我们对每个结果分配的数字：

阶段 1：

0 = 死亡

1, 2, 3, 4, 5, 6, 7, 8, 9 = 存活

阶段 2：

0, 1, 2, 3, 4, 5 = 移植成功

6, 7, 8, 9 = 仍需洗肾

阶段 3，有新肾：

0, 1, 2, 3, 4, 5, 6 = 存活五年

7, 8, 9 = 未能存活五年

阶段 4，洗肾：

0, 1, 2, 3, 4 = 存活五年

5, 6, 7, 8, 9 = 未能存活五年

第 3 阶段的数字分配，和第 2 阶段的结果有关。所以二者间不独立。

第 3 步：以下是好几次模拟的结果，每次的结果从上往下用一栏表示。我用了表 A 中列 110 的数字。



	第 1 次	第 2 次	第 3 次	第 4 次
阶段 1	3 存活	4 存活	8 存活	9 存活
阶段 2	8 洗肾	8 洗肾	7 洗肾	1 新肾
阶段 3	4 存活	4 存活	8 死亡	8 死亡

莫里斯在我们的 4 次模拟中，有 2 次存活超过五年。现在我们了解如何安排这项模拟之后，应该交给电脑去进行多次重复。经由许多次的模拟，或者经由数学计算，我们得知莫里斯的五年存活概率是 0.558。

网络寻奇

网络上有许多模拟各种随机现象的小程式。有个有趣的问题叫“布方之针”(Buffon's needle)：在一张纸上画许多条各相距一英寸的直线，然后拿一根一英寸长的针，让它从纸面上方掉落。针会和某条直线相交的概率是多少？你可以在伊利诺伊大学香槟分校的研究生里斯(George Reese)的网站上找到关于布方之针的数学解以及模拟，网址是 <http://www.mste.uiuc.edu/reese/buffon/buffon.html>。个人网站有时候会消失：搜寻“Buffon's needle”可找到其他网站。结果概率是 $2/\pi$ ， π 乘上圆的直径就是该圆的周长。所以这个模拟也可以用来当做计算 π 的一种方式，而 π 是数学中最有名的数之一。



本章重点摘要

如果我们知道每个结果的概率，就可以用随机数字来模拟随机结果。我们依据的事实是，每个随机数字是从 0 到 9 的 10 个可能数字之一，任一个的概率都是 0.1，以及随机数字表中的所有数字之间是互相独立的。如果要模拟更复杂的随机现象，可模拟各个阶段再串连起来。常常出现的状况是有好些个互相独立的试验，而每次试验的可能结果和概率都是相同的。想想掷铜板好几次或者把骰子掷好几次的情况。其他的模拟中，也许所需试验的次数不固定，或者每一阶段的概率不同，也或者有彼此之间不独立的各阶段，以致于其中某些阶段的概率和较早阶段的结果有关。成功模拟的关键，是先把概率模型仔细考虑清楚。



第19章 习题

19.1 哪一党的表现比较好？有一项民意调查随机选取美国成年人并问他们：“民主党和共和党这两党之中，你认为何者较善于处理经济问题？”针对以下每一个情况，仔细说明会如何从表 A 分配数字，来模拟个人的回答。

- (a) 所有美国成人当中，50% 会选择民主党，50% 选择共和党。
- (b) 所有美国成人当中，60% 会选择民主党，40% 选择共和党。
- (c) 所有美国成人当中，40% 会选择民主党，40% 选择共和党，另有 20% 无法决定。
- (d) 所有美国成人当中，53% 会选择民主党，47% 选择共和党。

19.2 小型意见调查。假设一所大学中有 80% 的学生赞成废止在晚间考试。你问了 10 位随机选择的学生，10 位都赞成废止晚间考试的概率是多少？

- (a) 独立的问 10 位学生的概率模型是什么？
- (b) 分配数字，分别代表“赞成”及“不赞成”。
- (c) 重复模拟 25 次，从表 A 的列 129 开始。你的估计概率是多少？

19.3 基本模拟。针对习题 19.1 的 4 种情况，利用表 A 分别模拟 10 位独立选出的成人的回应。

- (a) 针对情况(a)，用列 110。
- (b) 针对情况(b)，用列 111。
- (c) 针对情况(c)，用列 112。
- (d) 针对情况(d)，用列 113。

19.4 模拟意见调查。一项近期做的意见调查显示，已婚女性中约有 70% 认为，他们的先生做了至少分内该做的家务。假设这是完全正确的，则假如随机选择一位已婚女性，她认为老公有做足够家务的概率就是 0.7。如果我们个别访问一些女性，就可以假设她们的回答会是互相独立的。我们想要知道，一个 100 位女性的简单随机样本，会包含至少 80 位认为老公做了分内的家务的女性的概率。仔细说明这项模拟要怎么做，并且用表 A 的列 112 模拟一次调查。100 位女性



中有几位说老公做分内的家务?请说明怎样借助多次模拟来估计我们要找的概率。

19.5 修课成绩。从上沃巴什技术学院近年所有修过统计入门的学生当中,随机选出一位。这位学生在该科目所得成绩的概率如下:

成绩	A	B	C	D 或 F
概率	0.2	0.3	0.3	?

- (a) 他得 D 或 F 的概率一定是多少?
- (b) 若要模拟随机选择的学生的成绩,你会怎样分配数字,来代表列出来的 4 种可能结果?

19.6 班级排名。随意选一位大学生,问他在高中时的班级排名。各结果的概率如下:

	最高 25%		最高 50%	
班级排名	最高 10%	但非最高 10%	但非最高 25%	最低 50%
概率	0.3	0.3	0.3	?

- (a) 一个随机选择的学生,以前高中时在班上排名为后一半的概率是多少?
- (b) 若要模拟一个随机选择的学生的高中班级排名,你会怎样分配数字来代表列出的 4 种可能?

19.7 修课成绩续论。在习题 19.5 当中,你说明了怎样可以模拟随意选择的一个修统计课的学生们的成绩。宿舍里面同一层楼有 5 个学生正在修这门课。他们不一起读书,所以他们的成绩互相独立。利用模拟来估计,这 5 个人的修课成绩都有至少 C 以上的概率。(模拟 20 次。)

19.8 班级排名续论。在习题 19.6 当中,你说明了怎样可以模拟一个随机选择的大学生在高中时的班级排名。“随机基金会”决定要提供 8 位随机选择的学生全额奖学金。8 位随机选择的学生中,至多有 3 人高中时的班级排名在后一半的概率是多少?模拟该基金会的选择



10次，来估计这项概率。

19.9 沙奎尔的罚球。职业篮球队员沙奎尔·奥尼尔在整个球季中的罚球，差不多有一半会中。我们就把他每次罚球的投中概率当做是0.5。利用表A的列122，模拟他在一场球赛中罚12次球的表现，共模拟25回。这个模型就和掷12次铜板的模型一模一样。

- (a) 估计沙奎尔罚球12次会至少中8次的概率。
- (b) 检查你做的25回模拟当中，每一串投进及未投进的序列(sequence)，最多的连续进球数为多少？最多又有多少次连续未中？

19.10 唐雅的罚球。唐雅在一个长长的球季中，罚球命中率是70%。在一场比赛快结束的一段时间，她总共罚球5次，却有3次未进。球迷认为她太紧张了，但是投不进也可能完全是机遇巧合。我们来估计一下概率，把这件事弄清楚些。

- (a) 描述一下，如果投进一球的概率是0.7，怎样可以模拟1次罚球的结果。然后描述怎样可以模拟5次独立罚球的结果。
- (b) 以5次罚球为一回合，模拟50回合并记录每一回合未投进的球数。利用表A从列125开始。唐雅在5次罚球中，会有至少3次未中的近似概率是多少？

19.11 一考再考。伊莲修了一门可自己决定进度的课程，一共可以有3次考试机会来通过这门课程。她根本不读书，每次考试靠运气通过的概率是0.2。伊莲在3次考试机会中会通过的概率会是多少呢？(假设3次考试之间互相独立，因为每次考题不同。)

- (a) 说明怎样可用随机数字来模拟1次考试的结果。
- (b) 伊莲只要一通过某次考试就不必再考了。(这状况很像例3。)模拟50个回合，从表A的列120开始。你对伊莲会通过课程的估计概率是多少？
- (c) 你觉得假设伊莲每次考试的通过概率都一样合不合理？为什么？

19.12 一考再考的较佳模型。对于上一题当中伊莲试图通过考试的情况，以下为较合理的概率模型。第一次考时，她通过的概率为0.2。如果第一次没有考过，她在第二次通过的概率增加到0.3，因为考过一次总学到些东西。如果两次都没过，则第三次通过的概率是0.4。一旦她通过就不必再考试。但根据规定，不管有没有考过，顶



多只能考 3 次。

- (a) 把伊莲的考试过程用树图表示出来。要注意她在每一阶段的通过概率都不一样。
- (b) 说明怎样可以模拟伊莲试图通过课程的一回合考试。
- (c) 总共模拟 50 个回合，估计伊莲通过课程的概率。用表 A，从列 130 开始。

19.13 古罗马时代的赌博，掷 4 块距骨是古罗马时代最受欢迎的机遇游戏。把当今绵羊的距骨拿来掷许多次之后，显示出骨头在落地后会朝上的四个面之近似概率分布如下：

结果	概率
窄而平的一面	1/10
宽而凹的一面	4/10
宽而凸的一面	4/10
窄而凹的一面	1/10

掷 4 块距骨最好的结果叫“维纳斯”，这是朝上的四个面都不一样的情况。

- (a) 说明怎样可以模拟掷一块距骨的结果。然后说明怎样可以模拟掷 4 块距骨，且彼此之间互相独立的状况。
- (b) 模拟 25 回合掷 4 块距骨的结果。估计掷出“维纳斯”的概率。要写出你用的是表 A 的哪个部分。

19.14 亚洲随机甲虫 (Asian stochastic beetle)。我们可以利用模拟，来研究生物体的未来命运。考虑亚洲随机甲虫的情况。这种昆虫的雌虫有如下的繁殖模式：

- 20% 的虫在还没生雌幼虫之前就死掉，30% 生 1 只雌虫，50% 生 2 只雌虫。
- 个别雌虫的繁殖情况互相独立。

亚洲随机甲虫这个群体的前途会：繁殖很快？勉强保持数目？还是会渐渐灭绝？只要还存在一些雄虫，我们看雌虫的情况就足够了。

- (a) 分配数字来模拟 1 只雌甲虫的下一代。
- (b) 用树图画 1 只雌甲虫的雌性后代，总共画出三代。比如说第二



代可能有 0 只, 1 只或 2 只雌虫。如果是 0 只, 图就中止不再画下去。若不是 0, 则我们可模拟出每 1 只第二代雌虫的后代。三代之后的甲虫数目是多少?

- (c) 利用表 A 的列 105, 模拟 5 只甲虫的后代至第三代为止, 到第三代为止每只甲虫共有几个后代? 甲虫群体看来会增长吗?

19.15 两种警告系统。一架民航机有两套独立的自动系统, 在前方出现地形时(这是指飞机快要撞山了)会发出警告。两种系统都非常十全十美。系统 A 会及时警告的概率是 0.9, 系统 B 是 0.8。只要有一个系统正常运作, 驾驶员就会接到警告。

- (a) 说明如何模拟系统 A 对地形的反应。
(b) 说明如何模拟系统 B 的反应。
(c) 两种系统同时都在运作。画一个树图, 把系统 A 当作第一阶段, 系统 B 当作第二段阶段。模拟 100 个回合对前方地形的反应。估计会发出警告的概率。同时用两个系统, 其概率会高于只用 A 或只用 B。

19.16 骰子游戏。有一种游戏是用两个骰子玩的。参加的人掷两个骰子, 如果结果(两个骰子面上的点数和)是 7 或 11 他就赢了。若结果是 2、3 或者 12, 他就输了。假如是其他的结果, 他必须继续掷, 直到掷出和第一次一样的结果就赢了, 但若是掷出 7, 就输了。

- (a) 说明怎样模拟掷一个均匀骰子的结果。(提示: 只用 1—6 这些数字, 其他的不用理会。)然后说明怎样可以模拟那两颗骰子。
(b) 画出玩一次上述掷骰游戏的树图。理论上来说, 这个游戏可能永远玩不完, 不过你的图只要画到掷 4 次为止。用表 A, 从列 114 开始, 模拟玩这个游戏, 并且估计玩的人会赢的概率。

19.17 机场载客服务。你的公司经营从机场载客到城里旅馆的服务。一辆厢型车可载 7 位乘客。许多预约的乘客会不出现, 而且事实上, 任意一位乘客不出现的概率是 0.25。乘客之间互相独立。假如你每辆厢型车接受 9 位乘客预约, 结果会出现超过 7 位乘客的概率是多少? 用模拟来估计这个概率。

19.18 全是选择题的考试。迈特有过许多次没读什么书而参加选择题测验的经验。他就要考一个小考, 考题是 10 道选择题, 每题有 4



个答案。以下是迈特的个人概率模型。他认为在 60% 的题目当中，他会有办法消除一个一定不对的答案：然后他从剩下的 3 个答案当中猜 1 个。这样他猜中的概率是 $1/3$ 。另外的 40% 题目，他得从 4 个答案当中猜，猜中的概率是 $1/4$ 。

(a) 替一个题目的结果画树图。说明如何模拟迈特在一个题目上成功或失败。

(b) 题目之间互相独立。要模拟整个小考，只要模拟 10 个题目即可。迈特必须答对至少 5 题才能通过小考。你可以模拟很多次小考来找出他通过的概率，不过我只要求你模拟一次。请问迈特这次小考有没有通过？

19.19 机场载客问题续论。我们再继续 19.17 题的模拟。你有一辆备用厢型车，但是这辆车还要跑其他地方。在任意时间这辆车可以到机场来载客的概率是 0.6。你想知道，有些预约乘客会因为第一辆车已满，而第二辆又来不了，因而得滞留机场的概率。画一个树图，把第一辆车（不管有没有客满）当作第一阶段，第二辆车（不管能不能来）当作第二阶段。在习题 19.17 当中你模拟了第一阶段若干次。在每次第一辆车客满的情况下，都加入第二阶段的模拟。你估计会有旅客滞留的概率是多少？

19.20 生日问题。概率论里面有一个著名的例子，算出只要一间屋子里有 23 个人，则至少有两人同一天生日的概率就已经超过 $1/2$ 。概率模型如下：

- 随意选一个人，他在一年 365 天当中任一天出生的概率是一样的。
- 屋内不同的人之生日是独立的。

要模拟生日，必须让表 A 当中的每 3 个数字一组，代表一个人的生日。也就是说，001 代表元月 1 日，而 365 代表 12 月 31 日。忽略闰年这回事，也跳过不代表生日的其他 3 位数。用表 A 的列 139 来模拟随意挑选的人的生日，直到有同一个生日出现第二次时为止。你一共检视了多少个人，才找到两个同一天生日的人？

用电脑可以轻易重复这个模拟许多次。你可以找出 23 个人当中至少有两人同一天生日的概率；或者预期要问多少人，才会找到两个同一天生日的人。这些问题要用数学来算有点难，所以可显出模拟的优势与重要。



19.21 乘法规则。以下是另一个基本概率规则：如果几个事件之间互相独立，则所有事件都发生的概率，等于个别事件概率的乘积。比如说，我们知道生女孩的概率是 0.49，生男孩的概率是 0.51，而前后出生的孩子性别是独立的。所以一对夫妇的两个孩子都是女孩的概率为 $(0.49)(0.49) = 0.2401$ 。你可以用这个乘法规则来计算出我们在例 3 中模拟的概率。

- (a) 把 3 个孩子的所有 8 种性别组合可能都列出来，比如说，女女女和女女男。用乘法规则算出每种可能的概率。把你的 8 个概率加起来看是不是等于 1，来检查你算得对不对。
- (b) 例 3 当中的夫妇计划生到女孩就停，或者生到 3 个孩子时停，不管是不是全生男孩。用你在 (a) 中算出的结果，来计算出他们会生到女儿的概率。

第 20 章

赌场的优势：期望值

赌场的优势和赌场的真正优势

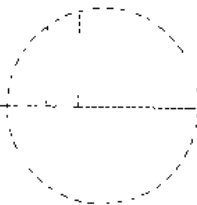
如果你赌博的话，你会在乎你是不是常赢。赢的概率告诉你，你赌许多次当中会赢的比例。你更在乎的是你会赢多少，因为赢很多比只赢一点点要好得多。如果你玩的游戏有 50% 机会赢得 1 美元，长期下来你有大约一半的时候会赢，因此平均每玩一次赢 50 美分。相较之下，玩有 10% 机会赢 100 美元的游戏会比较划算，因为虽然长期下来 10 次当中你只赢 1 次，但是平均起来你每玩一次可赢 10 美元。这些“每玩一次平均赢多少”的数字，就是期望值(expected value)。就和概率一样，期望值只告诉我们，赌很多次的时候，平均起来会是什么状况。



赌场会设定机遇游戏的期望值,使得长期下来平均来说赌场会赢,这可不是什么秘密。以轮盘为例,赌1美元的期望值只有0.947美元,平均起来每赌1美元有5.3美分进了赌场的口袋。因为对赌场来说,它每个月要玩上几十万次轮盘,所以已差不多可以确定,每1美元的赌注它可保有5.3美分。

但大部分赌客不知情的一点是,真实世界里的赌场,获益要比期望值好得多。事实上,赌场在轮盘上赚的钱,比赌客下注金额的20%还多一些。那是因为赢了钱的赌客还会继续赌。你就想想有这样一位赌客,他每次赌1美元可以赢回赌注的95%,赌完一次,他还有95美分;赌完两次,他有95美分的95%,也就是90.25美分;赌完三次,他有90.25美分的95%,即85.7美分。他玩得愈久,他原来的1美元就有更大的一部分变成赌场所有。真正在赌时当然并不是每赌一次就会得回固定的百分比,但是即使是最幸运的人,只要玩得够久,就一定会输。以期望值来说,每赌1美元赌场赚5.3美分,但是实际上,只要有1美元进了赌场的门,就会留下20美分给赌场。

这就让我们想到百家乐(baccarat)了,这是有钱有闲的人爱玩的游戏。百家乐是一种既不需做决定也不用任何技巧的纸牌游戏,但是看起来很优雅。有人一把就赌10万美金。下大赌注的人不多,而这些人不像玩较大众化的轮盘的人那样会玩很多把。所以“长期来说”赌场的获利,在百家乐上头并不丰厚。若有人在轮盘游戏上手气很好一再赢钱,赌场在1个月左右就可以把收支平衡过来,但这若发生在百家乐上,赌场可就平衡不过来了。因此有些赌场的获利季报,会因在百家乐赌台运气不好而受影响。



期望值

对机遇结果下注赌博,这种事从古时候就有了。在美国早期的时



候，公家和私人的彩券都很普遍。美国政府经营的赌博在消失了大约一世纪之后，于 1964 年重现江湖，新罕布什尔州提出一种彩券，以不加税的方法提高公共收益，造成了一股狂热。但在一些较大的州都采用这种概念以后，狂热迅速消退。现在共有美国的 37 个州以及加拿大所有的省发行乐透彩。州彩券因为把赌博变成娱乐而被接受。美国 50 个州当中，有 48 个州允许某种形式的合法赌博。所有美国成人当中，超过一半曾经合法赌博。他们花在赌博的钱，超过花在观赏运动比赛、玩电动游戏、主题公园和电影票的总和。如果你要赌的话，应该要了解怎样赌比较划算。就像本章刚开始的例子中所说的，我们不仅在意赢的概率，也在意赢多少。

例 1 三州的每日一数

以下是一种简单的乐透彩赌法，这是新罕布什尔、缅因和佛蒙特三州共同主办的“三州的每日一数”中，选 3 个数字名为“连胜”的游戏。你付 1 美元，选择一个 3 位数。州政府用随机方式选出一个 3 位数得奖号码，若选出的号码和你的一样，你就赢 500 美元。因为 3 位数共有 1 000 个，所以你赢的概率是 $1/1\,000$ 。以下是你赢多少钱的概率模型，单位是美元：

结果	0	500
概率	0.999	0.001

平均来说你赢多少？平常，两个可能结果 0 美元和 500 美元的平均会是 250 美元，不过把这当做平均所赢的数目毫无道理，因为得到 500 美元的机会比得到 0 美元的机会小得多。长期来说，每赌 1 000 次你才赢 1 次 500 美元，其他 999 次都会输。（当然如果你买整整 1 000 次选 3 个号码的彩券，也不保证你一定会赢一次。概率只是长期比例）你买一张彩券的长期平均所得是：

$$500 \cdot \frac{1}{1000} \text{ 美元} + 0 \cdot \frac{1}{1000} \text{ 美元} = 0.50 \text{ 美元}$$

即 50 美分。你看到了，长期来说州政府只把收进的赌注付出一半当彩金，而自己留了一半。



我们在例1中用来评估值不值得赌的这种“平均结果”(average outcome),有一个一般定义如下。

• 期望值

有数值结果的随机现象,其期望值(expected value)是每一个结果乘上它的概率,然后再把所有可能的结果加总而得。

如果用符号表示,假设可能结果是 a_1, a_2, \dots, a_k , 它们的概率是 p_1, p_2, \dots, p_k , 则期望值是:

$$\text{期望值} = a_1 p_1 + a_2 p_2 + \dots + a_k p_k$$

期望值是所有可能结果的平均,但是不像平常的平均一样把所有结果一视同仁。而是要把每个结果依照它的概率来加权,所以比较经常发生的结果,就有比较高的权重。

例2 再论三州的每日一数

例1中的连胜彩券,在你选的3位数和得胜数字完全相同时会付你彩金。你也可以选择另一种1美元的“连胜盒”。你还是要选一个3位数,但是现在有两种情况可能赢钱。如果你的3位数和得奖号码一样,你赢292美元;如果你的3个号码和得奖号码一样,但顺序不同,你就赢42美元。比如说,假设你的号码是123,则得奖号码是123时你赢292美元,而当得奖号码是132、213、231、312和321其中之一时,你就赢42美元。长期下来,每赌1000次你就有1次赢292美元,而有5次赢42美元。所以你赢得数目的概率模型为:

结果	0	42	292
概率	0.994	0.005	0.001

期望值为:

$$\begin{aligned} \text{期望值} &= (0 \text{ 美元})(0.994) + (42 \text{ 美元})(0.005) + (292 \text{ 美元})(0.001) \\ &= 0.502 \text{ 美元} \end{aligned}$$



在乐透彩上做手脚

美国人都在电视上看过乐透开奖的实况转播，看到号码球上下乱跳，然后由于空气压力而随机弹跳出来。怎么样可以对开出的号码做手脚呢？1980年只跳出3个号码球的时候，宾州乐透事实上被面带微笑的主持人以及几个舞台工作动了手脚。他们把10个号码中的8个号码球注入油漆，这样做会把球变重，因此可保证开出得奖号码的3个球必定是那2个没被注入油漆的号码。然后这些家伙就下注买该2个号码的所有组合。当6-6-6跳出来的时候，他们赢了120万美元。是的，他们后来被逮捕了。

看来连胜盒的赌法比连妥善处理稍稍好一点，因为州政府付出的彩金比一半赌注多一些。

“三州的每日一数”在所有州办彩券游戏当中，是比较特别的一种，它的各种赌法所付出的彩金是固定的。大部分的州，付彩金时是用“同注分彩赌博”(pari-mutuel)系统。新泽西州的“选三号”游戏就是一个典型：州政府把下注金额全部加起来，用一半来当作彩金，由所有得奖的彩券均分。你还是有1/1000的概率赢得连胜，但是你选的号码123会替你赢多少钱，一方面要看那天下注于“选三号”游戏的赌金有多少，另一方面要看有多少人选了这个号码。因为彩金不固定，今天赌123这个号码的期望值就没办法算，但是有一点是不变的，就是州政府把一半下注金额入袋。

把期望值当作平均的这种概念，不只可以用在机遇游戏，也可以用在其他的随机结果上。比如说，它也可以用在描述买股票或建新工厂等不确定回收上面。以下是不一样的一个例子。

例3 每户几辆车？

美国的住户，平均一户有几辆车？普查局告诉我们，每户车辆数的分布(1997年资料)如下：

车辆数	0	1	2	3	4	5
住户比例	0.04	0.25	0.45	0.18	0.06	0.02

这是任意选一个住户，并数数它有几辆车的概率模型。(有极少数住户有超过5辆车，我没有考虑他们。)这个模型的期望值，是一个住户的平均车辆数。这个平均是：

$$\begin{aligned}
 \text{期望值} &= (0)(0.04) + (1)(0.25) + (2)(0.45) + (3)(0.18) \\
 &\quad + (4)(0.06) + (5)(0.02) \\
 &= 2.03
 \end{aligned}$$



大数法则

期望值的定义是：它是可能结果的一种平均，但在计算平均时，概率大的结果占的比重较高。我认为期望值也是另一种意义的平均结果，它代表了如果我们重复赌很多次，或者随机选出很多住户，实际上会看到的长期平均。这并不只是直觉而已。数学家只要用概率的基本规则就可以证明，用概率模型算出来的期望值，真的就是“长期平均”。这个有名的事实叫做大数法则(law of large numbers)。

• 大数法则

大数法则(law of large numbers)是指，如果结果为数值的随机现象独立地重复许多次，实际观测到的结果其平均值会趋近期望值。

大数法则和概率的概念密切相关。在许多次独立的重复当中，每个可能结果的发生比例会接近它的概率，而所得到的平均结果就会接近期望值。这些事实表达了机遇事件的长期规律性。它们是真正的“平均数定律”。

大数法则解释了：为什么对个人来说是消遣或者嗜好的赌博，对赌场来说却是生意。经营赌场根本就不是在赌博。很大数量的客人平均赢的钱会很接近期望值。赌场经营者事先就算好了期望值，并且知道长期下来收入会是多少，所以并不需要在骰子里灌铅或者在洗牌时作弊来保证利润。

赌场只要花精神提供不贵的娱乐和便宜的交通工具，让顾客川流不息地进场就行了。只要赌注足够多，大数法则就能保证赌场赚钱。人寿保险公司的运作也很像赌场，它购买了保险的人不会死。当然，有些人确实会死，但是保险公司知道概率，并且依赖大数法则来预测必须给付的平均数目。然后保险公司就把保费订得足够高，来保证有利润。

深入探讨期望值

跟概率一样，期望值和大数法则都值得再花些时间，探讨相关的



高科技赌博

全美国有超过 450 000 台老虎机。从前，你丢硬币进去再拉一个把手转动三个轮子，每一个轮子有 20 个图案。但已经不再是这样子了。现在的机器是电动游戏，会闪出许多很炫目的画面，而结果是由随机数字产生器决定。机器可以同时接受许多硬币，有各种让你眼花缭乱的中奖结果，还可以联结起来产生共同满堂彩 (jackpot，会吐出机器里所有硬币)。赌徒仍在寻找可以赢钱的赌法，但是长期下来，随机数字产生器会保证赌场有 5% 的利润。

细节问题。

多大的数才算是“大数”？大数法则是说：当试验的次数愈来愈多时，许多次试验的实际平均结果会愈来愈接近期望值。可是大数法则没有说：需要多少次试验，才能保证平均结果会接近期望值。这点是要看随机结果的“变异性”决定。

结果的变异愈大，就需要愈多次的试验，来确保平均结果接近期望值。机遇游戏一定要变化大，才能保住赌客的兴趣。即使在赌场待上好几个钟头，结果也是无法预测的。结果变异性极大的赌博，例如累积奖金数额极大但极不可能中奖的州彩券，需要极多次的试验，几乎要多到不可能的次数，才能保证平均结果会接近期望值。(州政府可不需要依赖大数法则，因为乐透彩彩金不像赌场的游戏，乐透彩是用同注分彩赌博系统的。)

虽然大部分形式的赌博，变异没有彩券那么大，但对于大数法则的应用，以很实际的话来回答通常就是：赌场的参与次数够多，可以依赖大数法则，但你可不行。赌博能够诱惑人心，大部分是因为赌客对赌博的结果无法预测。而赌博这门生意依靠的则是：对赌场来说，结果并非不可测的。

有没有保证赢钱的赌法？把赌博很当回事的赌客常常遵循某种赌法，这种赌法每次下注的数目，是看前几次的结果而定。比如说，在赌输时，你可以每次把赌注加倍，直到你赢为止——或者，当然，到你输光为止。即使轮盘并没有记忆，这种玩法仍想利用你有记忆这件事来赢。

你可以用一套赌法来战胜概率吗？不行，数学家建立的另一种大数法则说：如果你没有无穷尽的赌本，则只要游戏的各次试验(比如输盘的各次转动)之间是独立的，你的平均获利(期望值)就会是一样的。很遗憾。

用模拟计算期望值

我们要怎么样实际计算期望值？你已经知道了数学公式，但是要用公式，必须先知道每个结果的概率。如果用这种方法计算期望值太困难，也可以用模拟方法来算。步骤还是跟以前一样：提出概率模型，用随机数字模拟，并重复许多次。根据大数法则，这些重复的平均结果，会接近期望值。

**例4 我们想要女儿，续集**

一对夫妇计划要生孩子直到生出女孩为止，或者生到3个孩子时停止，视哪种情况先发生而定。我们在第19章的例3，模拟了10次这个生孩子策略。并且估计了他们会有女儿的概率。

现在我们要问不一样的问题：采用这个计划的夫妇，平均来说会有几个孩子？也就是说，我们要算孩子数目的期望值。

模拟的方法和之前一样。概率模型说，前后生的孩子性别是独立的，而且每一个孩子是女孩的概率是0.49。

以下是我们较早做过的模拟结果，只不过现在不是记录那对夫妇有没有生女儿，而是记录他们生了几个孩子。提醒一下，我们是用两个数字模拟一个孩子，00—48(概率0.49)代表女孩。

6905	16	48	17	8717	40	9517	845340	648987	20
男女	女	女	女	男女	女	男女	男男女	男男男	女
2	1	1	1	2	1	2	3	3	1

这10次模拟的平均孩子数目是：

$$\bar{x} = \frac{2+1+1+1+2+1+2+3+3+1}{10} = \frac{17}{10} = 1.7$$

我们估计，如果许多夫妇照这个计划执行，他们平均会有1.7个孩子。这个模拟次数太少，结果不可靠。用数学计算，或者模拟很多次，会得到实际的期望值是1.77个孩子。



统计学上的争议|

合法赌博面面观

大部分人同意让某些方式的赌博合法，事实上也实行很多年了：乐透彩和赌场在美国和其他国家都很常见。赞成让合法赌博存在的人，用的理由很直接。很多人觉得赌博的娱乐性很高，也愿意花一点钱来换取一些刺激。而且赌博也不会伤害到别人，至少没有直接伤害。社会应该要容许受大部分人支持且无害的娱乐。州彩券还替教育之类的公益目的筹到钱，它有点像是不强迫人民缴纳的志愿税。

不过，反对赌博的人提出的有力论点，可能会使赌博受到更多限制。有的人发现赌博是会上瘾的。美国全国民意调查中心所做的一项研究，估计病态赌徒的贡献占有所有赌资的 15%，而这些人当中，每一个人终其一生会让纳税人花费 12 000 美元在社会及警察工作上面。赌博的确毁了某些人的一生，而且也间接对别人有害。

美国的州政府所办的彩券，还牵涉到由政府鼓励老百姓赌博的问题。在纽约州彩券发行之初，我记得见过广告看版上写着：“支持我们的教育——请买乐透彩。”结果没什么效果，广告很快就改成“一夜致富——请买乐透彩。”乐透彩通常只把下注金额的一半左右拿出来当彩金，所以如果要靠

赌博发财的话，乐透彩连赌场的老虎机都不如。职业赌徒和统计学家多半对乐透彩不屑一顾，因为浪费钱在这上面很不划算。穷人的收入当中花在乐透彩的比例比富人要高，而且穷人是“每日一数”游戏的主要参加者。乐透或许是一种志愿税，但是对穷人影响最大，而美国各州政府花费数以亿计的美元打广告，说服穷人去输掉更多钱。这里给个好心的建议：州政府应该少做点广告，并且把收益多拿些出来做彩金。

美国的州政府发执照给赌场，除了因为赌场付税并吸引游客外，当然也因为很多老百姓喜欢有赌场。事实上除了拉斯维加斯以外，大部分赌场所吸引的赌客，主要是住在附近的人。有赌场的县犯罪率较高，然而有许多的潜在变量可能可以解释这项关联。病态赌徒因犯罪被逮捕的比率的确较高，不过是否有因果关系仍然不确定。

开放赌博与否的争论持续不断。而与此同时，以网络赌博形式出现的科技已经绕过政府的监督，并创造了一个崭新的赌博环境，使得许多旧的论点都显是不合时宜了。



网络寻奇

有关合法赌博的争论持续不断。要看反对赌博的案例,可访问美国“全国反对合法赌博联盟”(National Coalition Against Legalized Gambling)的网站 www.ncalg.org。要看赌场业者的反驳,可访问“美国博弈协会”(American Gaming Association)的网站 www.americangaming.org。州彩券通过美国州省彩券协会(National Association of State and Provincial Lotteries)发表意见,网站是 www.naspl.org。美国的“全国印第安博弈协会”(National Indian Gaming Association)立场更坚定,网站是 www.indiangaming.org, 点击 Information 下的 Public Relations, 然后点击 Myths 可看到一些例子。由美国国会中研究赌博影响的委员会所做的报告,可以在下面网站看到, <http://govinfo.library.unt.edu/ngisc/index.html>。你在以上这些网站可以看到许多相关事实及数字。



本章重点摘要

概率告诉我们，随机现象每一个可能结果出现的频率(长期下来)。当结果是数字时，比如像机遇游戏的情形，我们也会想知道长期下来的平均结果会是什么样子。**大数法则**告诉我们，重复许多次之后的平均结果，迟早会靠近期望值。**期望值**是所有可能结果的加权平均，每个结果所对应的权重是该结果的概率。如果你不知道各结果的概率，你可以利用模拟来估计期望值(以及各结果的概率)。



第20章 习题

20.1 数字彩。选3个数字的游戏(例1)模仿自“数字彩”，数字彩是在大城市较穷的地区常见的地下赌博活动。其中一种版本的玩法如下：你在000—999之间的1000个3位数中任选1个，然后付1美元给当地的组头来下注。每天都会随机选出一组3位数，中了的人可以得600美元。赌一组号码的期望值是多少？对赌博的人来说，数字彩是否多少比例1中的选3个数字游戏划算？

20.2 选4个数字。“三州的每日一数”的选4个数字游戏，和例1的选3个数字很像。你付1美元，选一个4位数。州政府开出一个随机的4位数，若和你选的数字一样就付你500美元。赌1美元来玩选4个数字游戏的期望值是多少？

20.3 轮盘赌中的红色或黑色。美国的轮盘共有38格，其中18格黑色，18格红色，2格绿色。转动轮盘之后，球最后停在任何一格的概率是一样的。如果在红色上下注1美元，当球落在红色格子你会赢得2美元。(当赌客选择红色或黑色时，球若落入两格绿色格子就算是赌场赢。)写出对红色下注1美元会赢多少钱的概率模型，并找出这个赌注的期望值。

20.4 再谈选4个数字。就和选3个数字一样(例2)，选4个数字也有比较复杂的玩法。若你在连胜盒游戏下注1美元，而你选了1234，则若开奖时开出1234你赢2604美元，若开出1234这4个数字但是顺序不同，你就赢104美元。你赢钱金额的期望值是多少？

20.5 再论轮盘。赌客下注于轮盘时，是将筹码放在一张台子上，台面上列有轮盘的38格的号码及颜色。红色和黑色的格子，在台子上排成三行，每行有12格。如果在某一行上面赌1美元，当球落在该行12格中的任一格时你就赢3美元。这样玩所赢金额的期望值是多少？赌一整行的玩法对赌客来说，是否多少比赌黑色或红色要划算(习题20.3)？



20.6 做决定。心理学家特维斯基(Amos Tversky)以一般人对于机遇结果的看法为题,做了许多研究。《纽约时报》在特维斯基的计闻里引用了以下的例子。

- (a) 特维斯基要求受试对象在两种会影响 600 人的公共健康计划中选择其一。其中一项计划有 $1/2$ 的概率可以救全部 600 人,而有 $1/2$ 的概率全部 600 人都会死。另一计划保证在 600 人中可以救整整 400 人。算一算第一个计划可以救的人数的期望值。
- (b) 然后佛斯契又提供了另外的选择。一项计划有 $1/2$ 概率可以救全部 600 人,而且 $1/2$ 概率会损失全部 600 个人,第二项计划确定会损失共 200 条生命。这里的选择和(a)中的选择差别在哪里?
- (c) 要在(a)中做选择时,大部分受试者选择第二项计划。要在(b)中做选择时,大部分受试者选择第一项计划。看起来受试者有没有利用期望值来做决定?你觉得为什么在(a)和(b)两种情况的决定会不一样呢?

20.7 做决定。一个 6 面骰有 4 面是绿色,2 面是红色,骰子很均匀,每一面朝上的概率都一样。你必须在下列三种序列中选择一种:

红绿红红红

红绿红红红绿

绿红红红红红

现在开始掷骰子。假如掷出来的头几次结果和你选中的序列一样,你就可以赢 25 美元。

- (a) 哪一个序列的概率最高?为什么?(不必算概率你就可以看出来,哪个序列最有可能出现。)因为奖金 25 美元是固定的,所以概率最高的序列期望值就最高。
- (b) 在一项心理实验当中,260 位未学过概率的学生当中,有 63% 选择了第二个序列。请根据第 17 章“关于机遇结果的神话”当中的讨论,说明为什么大部分的学生没有选择最有机会赢的序列。

20.8 估计销售量。“盖恩通讯”销售飞机的通讯器材。下一年的销售量和市场状况有关,没有办法确实预测。盖恩遵循了当今用概率来估计销售量的做法。销售部经理预估下年度的销售量如下:



销售数量(套)	1 000	3 000	5 000	10 000
概率	0.1	0.3	0.4	0.2

这些是个人概率,代表销售部经理根据资料得出的个人意见。下一年度销售量的期望值是多少?

20.9 基诺(Keno)。基诺是赌场中很受欢迎的一种游戏。大家下注时,从1号到80号的号码球就在一个机器里面上下翻滚,然后从其中随机选出20个球。参加游戏的人在一张卡上标示出自己选的号码。以下是两种比较简单的基诺玩法。算出二种玩法赢钱金额期望值。

- (a) 在“标示一个号码”游戏下注1美元的话,如果你选的号码出现在随机选出的20个号码这中时你赢3美元;否则你的1美元就输掉了。
- (b) 在“标示2个号码”游戏下注1美元的话,如果你选的两个号码都出现在随机选出的20个号码之中时你赢12美元,这件事发生的概率大约是0.06。玩“标示2个号码”是不是多少要比“标示1个号码”划算些?

20.10 掷骰子。第18章的例2里面说到了掷两个骰子,并且记录朝上的两个面上的点数的概率模型。例子里也示范了如何计算点数和为5的概率。照着那个方法找出点数和概率模型。可能结果包括2、3、4……12。然后利用概率算出点数和的期望值。

20.11 亚洲随机甲虫。我们在习题19.14和这种昆虫见过面。雌虫的雌性后代数有以下的概率模型:

后代数	0	1	2
概率	0.2	0.3	0.5

- (a) 雌性后代的期望数目是多少?
- (b) 应用大数法则来说明,为什么雌性后代的期望个数大于1时,甲虫群体应会成长,而雌性后代期望个数小于1时,群体会渐渐灭绝。

20.12 期望值骗局?一位“未卜先知者”在杂志中登以下广告:

你怀孕了吗?知名占卜人士可以从母亲的任意一张相片,告诉你未出生的孩子性别。费用10美元。不准确包退费。



这可能是一种骗钱的把戏。假设占卜者对所有上门的人都说会生男孩，最糟的情况不过是有生女儿的人都来要回那 10 美元。把下面表中的空白处填入适当数字，并计算占卜者利润的期望值。

孩子性别	概率	占卜利润
男	0.51	
女	0.49	

20.13 再论亚洲随机甲虫 在习题 20.11 中你算出亚洲随机甲虫雌性后代的期望数目。模拟 100 只甲虫的雌性后代并算出这 100 只甲虫的平均后代数目。比较一下这个平均数和习题 20.11 算出的期望个数。(大数法则说，如果我们模拟的甲虫够多的话，平均数会很靠近期望个数。)

20.14 人寿保险。你考虑卖一种人寿保险给你一个 21 岁的朋友。21 岁男性在次年会死亡的概率大约是 0.001 5。你决定对一份若你朋友死亡会付 100 000 美元的保单，收取 200 美元保险费。

(a) 你这张保单的期望获利是多少？

(b) 虽然你预期利润不错，可是你卖这种保单给朋友就很笨了。为什么？

(c) 可以卖出几千份保单的人寿保险公司，若卖的保单条件和你的完全一样，会有不错的获利。说明为什么。

20.15 住户大小。普查局公开了以下美国住户人数的分布：

住户大小	1	2	3	4	5	6	7
比例	0.26	0.32	0.17	0.15	0.07	0.02	0.01

这也是随机选择一个住户，该住户人数多寡的概率分布。这个分布的期望值，就是住户的平均人数。这个期望值是多少？

20.16 修课成绩。一个大班统计课的成绩分布如下：

成绩	A	B	C	D	F
概率	0.2	0.3	0.3	0.1	0.1



要计算学生的学业平均成绩,先要把成绩等第用对应的数值表示,以 $A=4$ 、 $B=3$ 等等表示,一直到 $F=0$ 。

- (a) 求出期望值。这是这门课的平均成绩。
- (b) 说明一下如何可以模拟随机选取学生并记录他的成绩。模拟 50 名学生,并算出这 50 人的平均成绩。把这个估计的期望值和(a)中算出的确实期望值比较一下。(人数法则告诉我们,如果我们模拟非常多个学生的话,估计就会很准。)

20.17 我们真的很想要个女儿。例 4 中估计了一对夫妇的期望子女数,这对夫妇会生孩子生到有女儿为止,或生了 3 个就停止。假设他们不定上限,决定一直生,到生出女儿为止。这样子他们的期望子女数一定比例 4 当中的要高。你要怎样模拟这对夫妇的孩子状况?模拟 25 个回合。你估计的期望子女数是多少?

20.18 请来玩这个游戏。好了,朋友们,我这儿有个小游戏你可以玩。我们有一个很均匀的铜板(正、反面概率各 $1/2$)。掷两次。如果出现两个正面你就赢了。如果没出现两个正面,我就再给你一次机会;再把铜板掷两次,如果得到两个正面,你赢。(当然如果第二次试你还是没有掷出两个正面就我赢。)要玩的话你得付我 1 美元。如果你赢的话,我会把你的 1 美元还你,并且再给你 1 美元。

- (a) 替这个游戏画个树图。用树图来说明,怎样可以模拟玩这个游戏一次。
- (b) 你赌的 1 美元可能赢的金额有两种:如果我赢,你得 0 美元;如果你赢,你得 2 美元。用表 A 从列 125 开始,模拟 50 个回合。用你模拟的结果来估计你玩此游戏获利的期望值。

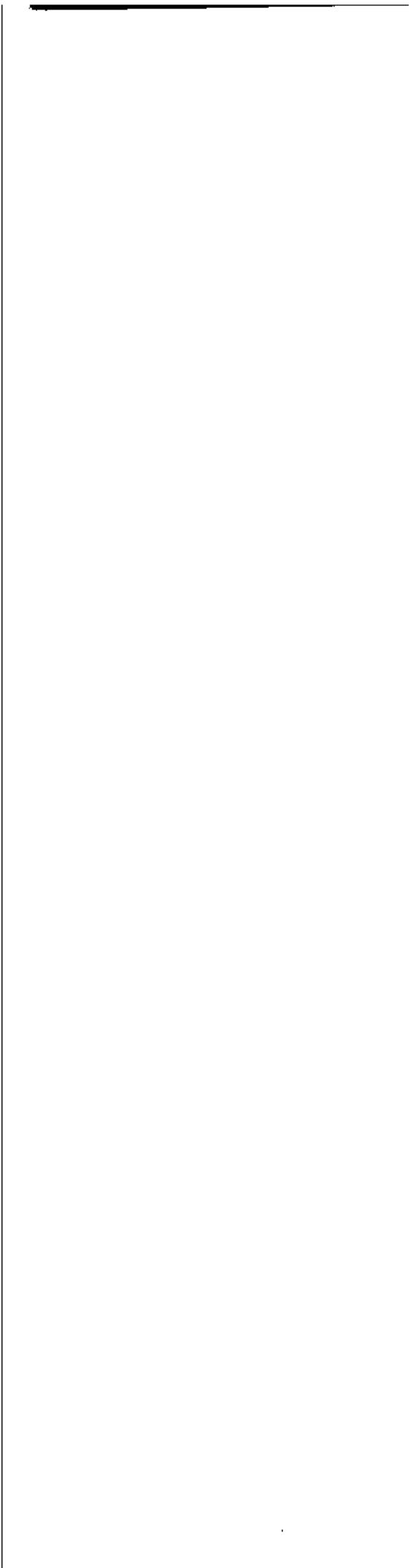
20.19 选择题测验。莎琳要考一个有 10 题选择题的小考,每题有 5 个答案。假设她每一题都独立的猜,则猜对一题的概率是 0.2。用模拟方法来估计莎琳答对题数的期望值。(模拟 20 回。)

20.20 一考再考。习题 19.12 中有某个自订进度的科目,至多可以考 3 次试以求通过的概率模型。在那题习题当中,你模拟了 50 个回合,用来估计伊莲通过课程的概率。用这个模拟结果(或重新做 50 个回合模拟亦可。)来估计伊莲考试次数的期望值。



20.21 共同期望值。以下是我们在 19 章模拟过的共同情境：次数固定的独立试验，每次试验的结果都是同样的两种可能和同样的概率。掷铜板、篮球赛中的罚球，以及观察新生婴儿之性别都是符合以上情境的例子。我们就把结果叫做“中了”或是“未中”。我们可以看出“中了”的次数期望值应该是多少。如果沙奎尔·奥尼尔罚 12 次球，而每次中的概率是 $1/2$ ，则罚中次数的期望值是 12 的 $1/2$ ，即 6。同样的道理，如果我们共有 n 次试验，而每次试验会中的概率是 p ，则中的次数的期望为 np 。这件事实可以用数学证明。我们不能用模拟来印证呢？模拟掷 10 次铜板 100 个回合。（要想快些做完的话，可以用表 A 总共 50 列当中每一列的头 10 个数及最后 10 个数，以奇数当作正面，偶数及 0 当作反面。）用 np 公式得到的期望正面个数是多少？你的 100 回合模拟的平均正面数是多少？

20.22 任意州的乐透彩。美国大部分的州都有一种乐透彩游戏，要从比如说 51 个号码中选 6 个，而头奖彩金都很高。任选一个有乐透彩的州，查一查销售金额中有多少百分比是拿出来当奖金的？多少百分比是州政府的营销费用？多少百分比是州政府的纯收益？州政府又把收益用在什么地方？



第三部分 复习

有些现象是随机的。虽然其个别单一结果事前无法预知，长期下来却有一种规则模式。赌博用具和抽取简单随机样本都是随机现象的例子。

概率和期望值提供了我们描述随机性的语言。随机现象的偶然程度与随机抽样相同。随机性其实是某种秩序，它有一种长期规律性，既非毫无章法，也不是在事前就已把事件固定的决定性机制。在第 17 章我们讨论随机性，第 18 章提出一些和概率相关的事实，而第 20 章讨论期望值。

当有随机性存在时，概率可以回答“长期下来多常发生”这样的问题，而期望值可以回答“长期下来平均是多少”这样的问题。

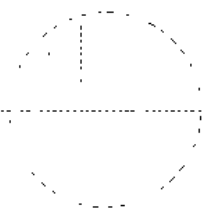
由于期望值用概率来定义，两个问题的答案因而息息相关。

“概率模型”对所有可能的结果分配概率，任何一个概率模型都必须符合概率规则。另外有种概率模型，用的是密度曲线，例如以正

态曲线底下的面积来分配概率。个人概率表达出对于某个事件有多大机会发生的个人判断。而个人概率如果要彼此相容，也必须遵守概率规则。

如果要计算一个较复杂事件的概率，又不要用数学来算，则可以用随机数字来模拟许多回合。期望值也可以用模拟方法来算。第 19 章教你如何模拟。先要有所有结果的概率模型，然后分配随机数字来模拟概率的分配。之后就可以用随机数字表来模拟许多回合。

把某一事件在多次模拟中发生的比例记录下来，就可以当作该事件概率的估计，而把平均结果记录下来就可以估计期望值。





第三部分 重点摘要

以下是你读完 17—20 章后，应该要具备的重要技能。

A. 随机和概率

1. 分得出有些现象是随机的。概率描述随机现象的长期规律性。
2. 了解事件的概率，是指某一随机现象重复许多次以后，该事件所发生的次数的比例。用“概率为长期比例”这个概念来思考概率问题。
3. 了解随机现象在短期之内并不见得会显示出概率所描述的规律性。相信随机现象在短期内是无法预测的，并且不会试图替随机发生的结果寻找可能的解释。

B. 概率模型

1. 能根据基本的概率事实，觉察到不合法的概率分配：任意概率都应该是在 0—1 之间的数，而且分配给所有可能结果的概率总和必定是 1。
2. 能根据基本的概率事实，求出由其他事件所形成事件的概率；一个事件不发生的概率，是 1 减去它发生的概率。如果两个事件不可能同时发生，则至少其中之一会发生的概率，是两个事件个别概率的和。
3. 分配概率时若分配给个别结果，则要找某一事件的概率时，就把组成该事件的个别结果的概率加起来。
4. 若概率是根据正态曲线来分配时，要找某一事件的概率时，就去找曲线底下的面积。

C. 期望值

1. 了解所谓的期望值，就是随机现象重复非常多次之后，所得数值结果的平均。
2. 懂得从列出所有可能结果和其概率的概率模型来算出期望值



(如果结果是数值的)。

D. 模拟

1. 能写出简单的概率模型来对每一个不同阶段分配概率，把各阶段当作彼此之间独立。
2. 会分配随机数字来模拟上述模型。
3. 会执行模拟许多回，来估计概率或期望值。



第三部分 复习习题

复习习题都是很短且直截了当的题目，能帮你加强在本书每一部中学到的基本观念及技巧。

III.1 概率是多少？把电话公司的住宅电话号码簿翻到任意一页。检视每个电话号码的最后4位。电话号码的前3位代表交换机，最后4位代表该交换机管辖范围内的个别号码。把你翻到的那一页前100个电话号码的倒数第4位记录下来。

- (a) 其中有多少个是1、2或3？电话号码的“个别号码”中的头一位为1、2或3的近似概率是多少？
- (b) 如果所有10个可能数字的概率一样，则得到1、2或3的概率应该是多少？根据你在(a)中做出的答案判断，你觉得电话号码的“个别号码”头一位是0—9当中任一数的概率是否相同？

III.2 修课成绩 从所有近几年修过数学101课程的学生当中随机选择一位。这位同学修课成绩的概率分布如下：

成绩	A	B	C	D	F
概率	0.2	0.3	0.3	0.1	?

- (a) 得到F的概率一定是多少？
- (b) 如果要模拟随机选择的一位同学的成绩，你会怎样分配数字来代表所列出的5种可能？

III.3 血型。随意选一个人，并记录他的血型。以下是每种血型的概率：

血型	O型	A型	B型	AB型
概率	0.4	0.3	0.2	?

- (a) AB型的概率必定为多少？
- (b) 若要模拟随机选择的人的血型，要怎样分配数字来代表四种血型？



III.4 修课成绩。如果你从习题 III.2 中所有修过课的学生中随机抽取 5 人, 其中每个人的成绩至少都是 C 的概率是多少? 模拟这个随机选取步骤共 10 个回合, 并利用你得到的结果来估计上述概率。(用 10 个回合结果做的估计当然不可靠, 不过只要你会模拟 10 个回合, 就会模拟 10 000 个回合。)

III.5 血型。血型 B 型的人可以接受 B 型或 O 型的人献血。泰若的血型是 B 型。泰若的 6 个好朋友中, 至少有 2 个人可以献血给她的概率是多少? 根据你对习题 III.3 的回答, 进行 10 个回合的模拟并估计此概率。(只根据 10 回合的结果做估计不可靠, 不过你已示范了原则上怎样可以找到这个概率。)

III.6 修课成绩。从习题 III.2 的课程中随机选取一个学生, 并观察该学生得到的成绩 ($A=4$, $B=3$, $C=2$, $D=1$, $F=0$)。

(a) 随机选择的一个学生, 其期望成绩是多少?

(b) 期望成绩结果指的并不是一个学生的 5 种可能成绩之一。说明一下为什么你的结果以期望值来说仍然是有意义的。

III.7 骰子。掷一颗均匀平衡的骰子一次, 朝上那面点数的期望值是什么?

III.8 风险投资的获利。某公司正在计划一项重要投资。可能获利并不确定, 但是概率估计可以用以下分布表示(单位为百万美元):

获利	1	1.5	2	4	10
概率	0.1	0.2	0.4	0.2	0.1

获利的期望值是多少?

III.9 扑克牌戏。用洗好的一副牌随意发出 5 张。这一手牌的几种可能概率大致如下:

手上的牌为	较差的牌	一对	两对	较好的牌
概率	0.50	0.42	0.05	?



- (a) 拿到的牌比两对还好的概率必定是多少?
- (b) 预期要等机把以后, 才第一次出现比“一对”更好的牌?说明怎样可以用模拟回答这个问题, 并且模拟两个回合。

III. 10 受了多少教育?《美国统计精粹》公布了随机选择的 25 岁以上美国人的教育程度分布:

教育程度	没有高中文凭	高中毕业	大专肄业	学士学位	更高学位
概率	0.17	0.34	0.25	0.16	0.08

- (a) 你怎么知道这是一个合法的概率模型?
- (b) 随机选择的一位 25 岁以上的人, 他至少有高中毕业学历的概率是多少?
- (c) 随机选择的一位 25 岁以上的人, 他至少有学士学位的概率是多少?

III. 11 学习语言。从美国某处的 9—12 年级学生当中随机选取一位, 并问他是否正在修习英文以外的语言。以下是结果的分布:

语言	西班牙语	法语	德语	其他语言	无
概率	0.26	0.09	0.03	0.03	0.59

- (a) 说明为什么这是一个合法的概率模型。
- (b) 随机选取的一名学生, 他正在修习英文以外语言的概率是多少?
- (c) 随机选取的一名学生, 他正在修习法语、德语或西班牙语的的概率是多少?

III. 12 随机选择。艾比、黛博拉、美琳、山姆及若柏托在某公司的公关部门工作。他们的雇主要在这 5 人当中选两人去巴黎开会。为了避免选择方式不公平, 因此将从帽子里抽出两个名子。(这是大小为 2 的 SRS。)

- (a) 写下从 5 个名字中选 2 个的所有可能选择。这些就是所有可能结果。
- (b) 随机选择使得每个结果的机会一样大。每个结果的概率是多少?
- (c) 美琳被选中的概率是多少?

(d) 两位男士(山姆及若柏托)都被选中的概率有多大?

III. 13 对大学的满意度。随意选择一位成年人,问他以下的问题:

“你所住的州里面,大专院校是办得非常好、好、尚可、差劲还是你的资料不足无法判断?”以下是意见的分布状况:

意见	非常好	好	尚可	差劲	无法判断
概率	0.15	0.42	0.13	0.03	?

(a) 随意选择一位成人,他的回答是“无法判断”的概率是多少?

(b) 随意选择一位成人,他认为大学办得好或非常好的概率是多少?

III. 14 IQ 测验。韦氏成人智力量表(WAIS)是一种普遍使用的成人 IQ 测验。16 岁以上的人之 WAIS 分数分布,大约是平均数 100、标准差 15 的正态分布。利用 68-95-99.7 规则来回答以下问题。

(a) 随机选择的一个人,他的 WAIS 分数在 115 以上的概率是多少?

(b) 所有 16 岁以上的人,中间 95% 的分数会在什么范围?

III. 15 我们喜欢民意调查。美国人对于有关当前重要议题的民意调查感兴趣吗?假设所有成人中有 40% 对这类民意调查非常感兴趣。

(从调查了这个问题的民意调查结果得知,40% 差不多是正确数字。)某民意调查公司选了一个 1 015 人的 SRS,如果这家公司这样子抽样很多次,样本中说非常感兴趣的比例会随样本而变,但遵循平均数 40%、标准差 0.5% 的正态分布。利用 68-95-99.7 规则回答下列问题。

(a) 以这样子抽样的一个样本,其结果会在总体真正值 $\pm 1.5\%$ 范围内的概率是多少?

(b) 以这样子抽样的一个样本,其结果会在总体真正值 $\pm 3\%$ 范围内的概率是多少?

III. 16 IQ 测验(此题可略过)。用习题 III. 14 中的信息和表 B,找出随机选择的一个人,其 WAIS 分数在 112 以上的概率。

III. 17 我们喜欢民意调查(此题可略过)。用习题 III. 15 中的信息和表 B,找出一个样本结果距离总体真正值至少有 4% 的概率。(这是



说样本结果要不是小于 36%，就是大于 44% 的概率。)

Ⅲ.18 IQ 测验(此题可略过)。在 WAIS 测试中，至少要得到多少分，才能列名在所有分数的前 10%?用习题Ⅲ.14 的信息和表 B 来回答这个问题。

Ⅲ.19 合法及不合法模型。一副桥牌有 52 张牌，其中有 A、K、Q、J、10、9……一直到 2，每种“面值”各有 4 张。从这样一副牌中任意发出一张，并记录下它的值。写出在整副牌洗得很彻底的情况下，各种结果应该有的概率。再分配一组和前面一组不同的合法概率(符合概率规则)。然后再分配一组不合法概率，并说明是哪里不合法。

Ⅲ.20 孟德尔的豌豆。孟德尔(Gregor Mendel, 1822–1884)用豌豆做了一些实验，实验结果显示，遗传是随机发生的。孟德尔用的豌豆种子，颜色有绿有黄。假设我们的种子是由两株植物杂交而来，二者都具有绿(绿色)和黄(黄色)两种基因。每株母株把黄和绿基因传给种子的概率约各为 $1/2$ ，母株和母株之间互相独立。除非两株母株都传下绿基因，否则种子均为黄色；而两株母株都传下绿基因的种子才会是绿色。这样杂交出来的种子，绿色出现的概率是多少?写出针对此问题的模拟过程，并模拟 25 个回合来估计概率。

Ⅲ.21 预测赢的队。“十大运动联盟”(Big Ten athletic conference)中共有 11 个队伍。以下是次年篮球冠军的一组个人概率：密歇根州立大学赢的概率是 0.3。而艾奥瓦、明尼苏达、西北及宾州州立大学一点机会都没有。这样还剩下 6 队。密歇根大学，俄亥俄州立大学及普渡大学赢的机会一样大。伊利诺伊、印第安纳和威斯康星大学赢的机会也一样，但是只有前面 3 所大学的一半。如此 11 队中每一队赢的概率各是多少?

Ⅲ.22 卖车。比尔靠卖新车生活。在工作日的下午，他会有一位顾客上门的概率是 0.2，两位顾客是 0.4，3 位顾客是 0.4。每位上门的顾客会买车的概率是 0.2。顾客和顾客之间，会不会买车是互相独立的。描述一下你会如何模拟比尔一个下午卖几辆车。你必须先模拟上门顾客的人数，然后模拟 1 位、2 位或 3 位顾客的购车决定。模拟一个回合来示范你的模拟过程。



第三部分 报告作业

报告作业是比较长的习题，需要搜集信息或制作数据，而且重点是要把做出的结果用一篇短文来说明。这里很多题目适合由一组学生共同来做。

作业 1. 来点历史。在第 17 章中我曾说：“对于随机的有系统研究，是从 17 世纪时法国赌徒请法国数学家帮忙算出机遇游戏的‘公开’赌注时开始。”费马和帕斯卡是提出回应的数学家中的两位。这两人都是很有趣的人物，请从中选择一位，写一篇关于他的短文，包括他何年出生何年死亡，一生中有哪些值得一提的轶事，且对他所研究的概率问题要至少提出一个例子。（以他们的姓名上网搜寻，就能找到许多信息。记得要用你自己的话来写短文。）



作业 2. 对风险的反应。在第 20 章中我引用了一位作者所说的：“我们就算只开车出去办 10 分钟的事，也很少有人会把婴儿独自留在家睡觉。”如果事实上，婴儿会在车中受伤的概率，的确远大于同时段会在家里受伤的概率。你会把婴儿单独留在家里吗？用一篇短文来说明你的理由。如果你不会把婴儿留在家里，别忘了必须说明为什么你不理会那些概率。

作业 3. 第一个数字。这是个值得注意的事实：较大的数字表里面，各个数字的第一码，会是 0、1、2、3、4、5、6、7、8、9，这 10 个数字中的哪一个，概率并不见得都一样。数字 1 出现的概率大约是 0.3，2 出现的概率大约是 0.17 等等。这个事实叫做本佛定律 (Benford's law)，你可以在网络上找到许多相关信息，也可参考以下两篇文章：希尔 (Theodore P. Hill) 写的 *The difficulty of faking data*，刊登在 *Chance*, 12(1999), No. 3, pp. 27–31, 以及 *The first digit phenomenon*，刊登在 *American Scientist*, 86(1998), pp. 358–363。要做这项作业并不是非读这两篇文章不可。

找出有很多数字的表至少两个，里面的数字应该是由什么数字开始都可能的。你可以选择数据表，比如很多城市的人口数，或者纽约证券交易所好几天当中的交易股数，或者诸如对数表、平方根表等数



学表格。我希望你很清楚不应该用随机数字表。我们就要求你的每个例子都至少包含 300 个数好了，把表里面每个数的第一码记录下来，对每个表都如此做。写出分布(用百分比)，比较不同的表产生的结果，并和班佛定律及相同概率分布做比较。

作业 4. 个人概率。个人概率是个人意见，所以应该会因人而异。选一件你的学校里大部分同学都会有意见的事，比如认为下周五会不会下雨，或者学校球队下一场会不会赢。询问许多学生(至少 50 人)，请他们告诉你，他们会分配给下雨或赢球什么概率。然后用图和数字来分析这些数据，包括形状、中心、离度等等。对于这个未来事件的个人概率，你的数据显示出了什么信息？

作业 5. 做决定。习题 20.6 里报告了心理学家特维尔斯基的一项研究结果，研究是关于措辞如何影响人们对机遇结果所做的决定。他的受试对象是大学生。在你的大学重做一次特维尔斯基的研究。准备两张打好的卡片一张上面说：

你负责治疗 600 位曾暴露于致命病毒的人。疗法 A 有 $1/2$ 概率可以救全部 600 个人， $1/2$ 概率全部 600 人都会死。疗法 B 保证可以救 600 人中的 400 人。你会用哪一种疗法？

第二张卡上说：

你负责治疗 600 位曾暴露于致命病毒的人。疗法 A 有 $1/2$ 概率可以救全部 600 个人， $1/2$ 概率全部 600 人都会死。疗法 B 确定会让 200 人丧失性命。你会用哪一种疗法？

把每一张卡都给至少 25 个人看(两张要分别给不同的 25 个人看，选人要在可能范围内尽量随机，而且要选没学过概率的人)。把他们的选择记录下来。特维尔斯基宣称看到第一张卡的人偏向选择 B，而看到第二张卡的人偏向选择 A。你的结果是否和他所宣称的一样？把你的发现做个简短的摘要；人们做决定时有没有考虑期望值？不同决定的出现形式(比如措辞)是否会影响选择？你还可以在下面这本书中找到许多类似奇怪想法的背景，作者：季洛维奇(Thomas Gilovich)，书名：《谬误辨析》(*How We Know What Isn't So: The Fallibility of Everyday Life*)Free Press 出版，1991 年。

第四部分

推论

推论的意思是根据证据做出结论。统计推论(statistical inference)是根据样本所提供的证据，对总体做出结论。在数学领域做结论，是要从某些假设开始，然后根据逻辑推理，证明结论确实毫无疑问绝对成立。统计却不一样。统计结论不是百分之百确定的，因为样本不等于整个总体。所以统计推论除了结论之外，还得说明结论的不确定程度。我们用概率语言来表示不确定的程度。

因为推论必须做出结论，并陈述不确定的程度，所以是统计里面最专业的部分。目的在训练人实际做统计的教科书和课程，大部分时间都花在推论上面。我这本书的目标，是帮助你了解统计，技巧不需那么多，思考却不可少。我们只会谈到推论的几个基本技巧。技巧很简单，背后的概念却很精妙，所以要准备开始思考了。首先想想你已学到的东西，不要被精巧的统计技巧给震慑住了：即使是用到最高度技巧的推论，也没法弥补诸如自发性回应样本或没有控制组的实验所产生的基本瑕疵。

第 21 章

什么是置信区间

别生气

你认识很容易生气的人吗？大自然有办法让这些人大平静下来：他们比较容易得心脏病。好几项观测研究都发现生气和心脏病之间的相关性。最好的一项研究观察了 12 986 个人，黑人白人都有，随机选自四个社区。首次检查时，所有受试对象的年龄在 45—64 岁之间，而且都没有心脏病。我们就把焦点集中在这个样本当中血压正常的 8 474 个人身上。

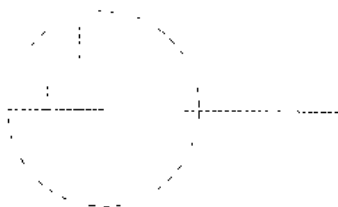
有个简短的心理测验叫做“斯皮尔伯格发怒量表” (Spielberger Trait Anger Scale)，度量了每个人容易发怒的程度。结果有 633 个人被归类在发怒量表的高阶，4 731 人在中阶，3 110 人在低阶。



然后追踪这些人将近 6 年并比较了高阶组和低阶组得心脏病的比率。有一些潜在变量存在：高阶组的人中较有可能是男性、高中没毕业以及既抽烟又喝酒的人。经过对这些差异做调整之后，最爱生气的高阶组和最不爱生气的低阶组比起来，得心脏病的机会是 2.2 倍，而心脏病猝发的概率是 2.7 倍。

听起来生气似乎是很严重的事。但是在研究期间，低阶组只有 53 个人，高阶组只有 27 个人得了心脏病。我们知道 2.2 倍和 2.7 倍这两个数字，对于所有正常血压的 45—64 岁人士来说不会是完全正确的。有关这项研究的新闻报道引用了这两个数字，但是医学期刊《循环》(Circulation)中的全文中提出了置信区间(confidence interval)。在 95% 置信水平之下，高阶组得心脏病的概率，是低阶组的 1.36—3.55 倍之间，而心脏病猝发概率则在 1.48—4.90 倍之间。区间提醒我们，因为我们只有样本的数据，所以我们对于总体的所有叙述都是不确定的。对样本来说，我们可以说：“机会恰好是 2.2 倍。”对整个总体来说，样本数据只能让我们说：“机会在 1.36—3.55 倍之间”，而且只有 95% 的信心。

要探讨新闻背后的真相，不论是有关医药行业还是其他领域，我们都必须使用置信区间这种表示方法。



估计

统计推论根据样本数据来对总体做结论，有一类结论是要回答“职业女性中有大专学历的占多少百分比？”或“得这类癌症的病人的平均存活时间是多少？”这类问题。这些问题所问的，是用来描述总体的一个数(百分比或平均数)。用来描述总体的数叫做**参数**(parameter)。要估计总体参数的话，我们从总体取一个样本，并利用从样本算出来的某个**统计量**(statistic)的值来当作我们的估计。下面是一个例子。



例 1 艾滋年代的危险行为

会有艾滋病风险的行为到底多普遍?美国的“全国艾滋行为调查”(National AIDS Behavioral Surveys)访问了 2 673 位成人异性恋者的随机样本。其中有 170 个人承认,在前一年曾有超过一个的性伴侣,占样本的 6.36%。这个结果可能有偏差,因为有人不愿意把自己的性行为照实告诉别人(参考习题 21.10)。目前就假设样本里的人都说了实话。根据这些数据,我们对于所有成年异性恋者当中,有不止一个性伴侣者所占的百分比,能下什么样的结论?

我们的总体是成年异性恋者。参数是在前一年中有不止一个性伴侣者所占比例。我们把这个未知参数叫 p , 代表比例(proportion)。用来估计参数 p 的统计量是样本比例(sample proportion) \hat{p} :

$$\hat{p} = \frac{\text{样本中的计数}}{\text{样本大小}} = \frac{170}{2\,673} = 0.063\,6$$

统计推论中的一个基本步骤,就是用样本统计量来估计总体参数。一旦我们取得样本之后,就可以估计所有成年异性恋者中有不止一个性伴侣的比例是“大约 6.36%”,因为样本里的比例正是 6.36%。我们只能估计总体的真正情况“大约”是 6.36%,因为我们知道样本结果通常不会和总体的真正比例一模一样。置信区间把这个“大约”具体化了。

95% 置信区间

95% 置信区间(95% confidence interval)是从样本数据计算出来的一个区间,保证在所有样本当中,有 95% 会把真正的总体参数包含在区间之中。

我们会先直接切进总体参数的置信区间,然后再讨论我们实际上做了什么,并且稍加推广。



有信心的估计

我们要估计总体成员中拥有某种特征的比例 p ，这个特征可能是他们有工作，或者他们对总统的表现满意，等等。让我们把正在考虑的这个特征叫做“成功”。我们会用简单随机样本(SRS)中的成功比例 \hat{p} ，来估计总体中的成功比例 p 。统计量 \hat{p} 做为参数 p 的估计，表现如何？要知道答案，我们会问：“如果我们取许许多多样本，会发生什么情况？”首先我们知道， \hat{p} 的值会随样本而变；我们也知道这个抽样变异并不是偶发的。长期下来它有很清楚的形态，用正态曲线可以把这个形态描绘得相当接近。以下就是相关事实。

• 样本比例的抽样分布

一个统计量的**抽样分布**，是指同一总体所抽出，同样大小的所有可能样本，其统计量的值之分布。从一个成功比例为 p 的很大总体抽取一个大小为 n 的 SRS。用 \hat{p} 表示成功的**样本比例**：

$$\hat{p} = \frac{\text{样本中的成功计数}}{n}$$

则当样本够大时：

- \hat{p} 的分布为近似正态(**approximately normal**)。
- 抽样分布的平均数和 p 相等。
- 抽样分布的标准差是：

$$\sqrt{\frac{p(1-p)}{n}}$$

这些事实是可以数学证明的，所以从这里出发，基础很坚实。

图 21.1 把这些事实用某种形式做了综合，也提醒了我们：抽样分布是在描述从同一总体抽出的许多样本的结果。

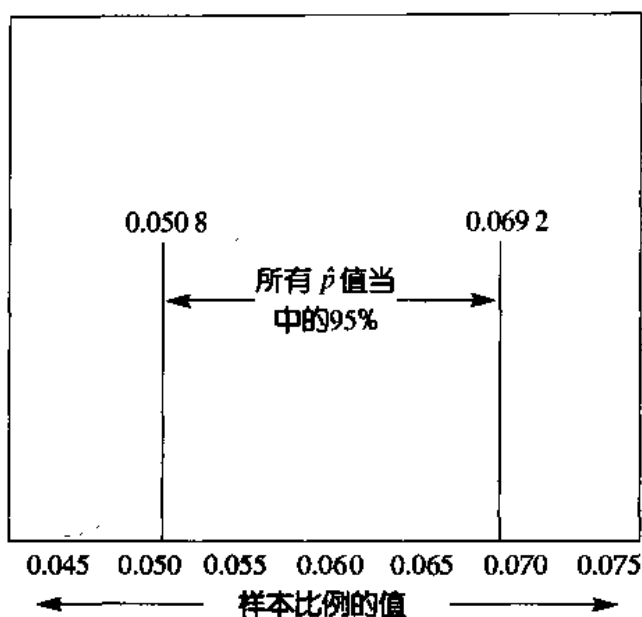


图 21.2 从成功比例 $p = 0.06$ 的总体，重复抽取大小为 2 673 的 SRS 许多次。样本比例 \hat{p} 最中间 95% 的值，会落在 0.050 8—0.069 2 之间

到目前为止，我们只不过是用数字把我们已经知道的事表达出来；我们可以信任大的随机样本的结果，因为几乎所有这类样本的结果，都很接近总体的真正值。数字告诉我们，大小为 2 673 的所有样本中的 95%，其统计量 \hat{p} 和参数 p 的值差距在 0.009 2 之内。也可以换个方式说：所有样本中有 95%，所得到的 \hat{p} 值会把总体真正值 p 夹在 $p - 0.009\ 2$ — $p + 0.009\ 2$ 之间。

而 0.009 2 是把 $p = 0.006$ 代进 \hat{p} 的标准差公式里得来的。对于任意 p 来说，一般性的事实如下：

当总体比例的值为 p 时，有 95% 的样本，其所得 \hat{p} 值往左右各延伸 2 个标准差所得到的区间，会把 p 值包含进去。

上面说的区间是：

$$\hat{p} \pm 2 \sqrt{\frac{p(1-p)}{n}}$$

这是不是我们要的 95% 置信区间呢？还不能说是。这个区间没有办法根据样本数据算出来，因为标准差公式里有总体比例 p ，而实际



上我们不知道 p 的值。在例 2 里我们把 $p=0.06$ 代进公式里，但这不见得是 p 的真正值。

该怎么办呢？是这样的，统计量 \hat{p} 的标准差的确是出 p 值决定，然而当 p 值改变时，标准差的值并不会改变太多。我们回到例 2，并重新计算对应其他 p 值的标准差。

算出的结果如下：

p 值	0.04	0.05	0.06	0.07	0.08
标准差	0.003 8	0.004 2	0.004 6	0.004 9	0.005 2

从这个结果可以看出来，如果我们猜测的 p 值合理靠近真正 p 值的话，用猜测的值算出来的标准差就会大致正确。我们知道当我们取的样本很大时，统计量 \hat{p} 的值几乎总是很靠近参数 p 的值。所以我们可以就用 \hat{p} 值当作我们猜测的 p 值。这样我们就有了一个可以根据样本数据算出来的区间。

• 比例的 95% 置信区间

从一个成功比例 p 未知的大总体抽取一个大小为 n 的 SRS。把这个样本中的成功比例叫做 \hat{p} 。参数 p 的一个近似 95% 置信区间为：

$$\hat{p} \pm 2 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

例 3 危险行为的置信区间

“全国艾滋行为调查”的 2 673 位成年异性恋者的随机样本中，有 170 人在前一年中有不止一个性伴侣，样本比例为 $\hat{p}=0.063 6$ 。所有成年异性恋者中，有不止一个性伴侣的人所占比例，其 95% 置信区间为：



$$\begin{aligned}
 \hat{p} \pm 2 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= 0.0636 \pm 2 \sqrt{\frac{(0.0636)(0.9364)}{2673}} \\
 &= 0.0636 \pm (2)(0.0047) \\
 &= 0.0636 \pm 0.0094 \\
 &= 0.0542 - 0.0730
 \end{aligned}$$

这个结果的解释如下：我们得到这个区间所用的方法，若用在所有的样本上面，会有 95% 的样本“抓到”未知的真正总体比例。精简来说就变成我们有 **95% 信心** 有个性伴侣的异性恋者真正比例，会落在 5.42%—7.30% 之间。

了解置信区间

总体比例的 95% 置信区间有我们熟悉的形式：

估计值 \pm 误差界限

我们知道对于抽样调查所做的新闻报道，举例来说，通常都把估计值和误差界限分开来报告：“根据一项最新的盖洛普调查，有 66% 的女性赞成设立新法规来对枪设限制。误差界限是正或负 4 个百分点。”我们也知道新闻报道常把置信水平省略不报。

并不是所有的置信区间都是这种形式。例如本章开头谈到的，爱生气的人有较大的心脏病发作的风险，其增加的风险的置信区间，形式就不一样。以下是置信区间的完整描述。

• 置信区间

一个参数的水平 **C 置信区间** (level C confidence interval) 有两部分：

- 一个由数据计算出来的区间。
- **置信水平 C** (confidence level C)，是不断重复抽样时，区间会抓到真正参数值的概率。



谁抽烟?

要估计一个比例 p 的时候, 要确知“成功”是指什么。新闻报道说青少年有 20% 吸烟。真是令人震惊。结果这个比例是指前 1 个月至少抽过 1 次烟的百分比。如果我们把吸烟者定义为过去 30 天中至少有 20 天有抽烟, 且在有抽烟的那天至少抽半包烟, 则青少年中的吸烟者还不到 4%。

置信区间公式有许多种, 可在各种不同的情况之下使用。要确实了解置信区间应如何解释, 而不管是哪个公式, 解释的方法都是一样的, 而且你没办法叫电脑来替你做这件事。

置信区间应用了概率的中心概念: 如果重复抽样许多次, 考虑会发生什么情况。95% 置信区间中的 95% 是概率, 是这个方法所产生的区间会抓到真正参数值的概率。

例 4 置信区间的面貌

全国艾滋行为调查的 2 673 位异性恋者样本当中, 有 170 个人有 2 个以上的性伴侣, 所以样本比例是:

$$\hat{p} = \frac{170}{2\,673} = 0.063\,6$$

而 95% 置信区间是:

$$\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.063\,6 \pm 0.009\,4$$

从同一总体再抽第二个样本。2 673 个人当中有 148 人有多个性伴侣。对应于这个样本:

$$\hat{p} = \frac{148}{2\,673} = 0.055\,4$$

$$\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.055\,4 \pm 0.008\,8$$

再抽一个样本。这回计数是 152, 因此样本比例及置信区间分别为:

$$\hat{p} = \frac{148}{2\,673} = 0.055\,4$$

$$\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.056\,9 \pm 0.009\,0$$

继续不断抽样下去。每个样本会产生一个新的 \hat{p} 和新的置信区间。如果我们如此不停地抽样下去, 所有的区间中有 95% 会包含真正的参数值。不论真正的参数值是什



么，都会是这样。

图 21.3 把置信区间的面貌，用图综合起来。

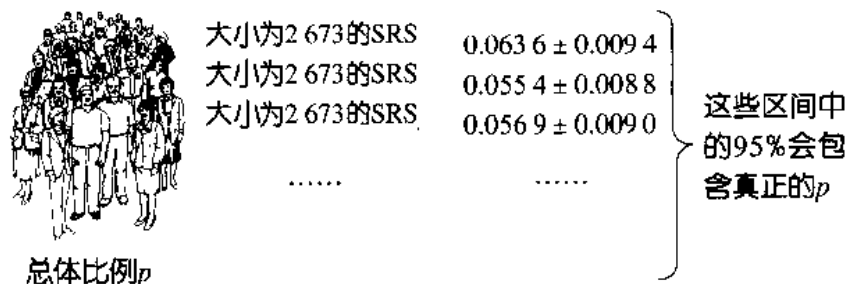


图 21.3 从同一总体重复抽样，会得到不一样的 95% 置信区间，但是这些区间当中，有 95% 会抓到真正的总体比例 p

因为两张图比一张图好，我们在图 21.4 里就从另外一个角度来看置信区间的面貌。例 4 和图 21.3 强调的是，重复抽样的结果会不同，而我们只能够确认，95% 的样本会产生正确结果。图 21.4 是从问题的背后去探讨。垂直直线代表总体比例的真正值 p 。图上方的正态曲线是样本统计量 \hat{p} 的抽样分布，中心点就在真正的 p 的位置。我说这是到问题的背后探讨，是因为在真实世界做统计时，我们是不知道 p 值的。

从 25 个 SRS 所得到的 95% 置信区间，一个接一个画在正态曲线下方。区间中的黑点代表 \hat{p} 值，位于区间的正中央。点两边的箭头一直延伸到区间的两端。长期下来，所有区间中有 95% 会涵盖到真正的 p 值，而有 5% 会漏掉。图 21.4 的 25 个区间当中，有 24 个包含了参数值，一个没包含。（要记得概率描述的是长期下来的情况，因此我们不能期望 25 个区间中，恰恰有 95% 包含真正的参数值。）

别忘记我们的区间只是近似的 95% 置信区间，不是确实的 95% 置信区间。不是确实的 95% 置信区间有两个原因：样本比例 \hat{p} 的抽样分布并不是确实符合正态分布，还有我们所用的 \hat{p} 值的标准差也不是完全正确，因为我们在公式中用 \hat{p} 取代了未知的 p 。这两个弱点造成的影响，会随着样本大小 n 的增加而愈来愈小。所以我们的公式只适用于较大的样本。

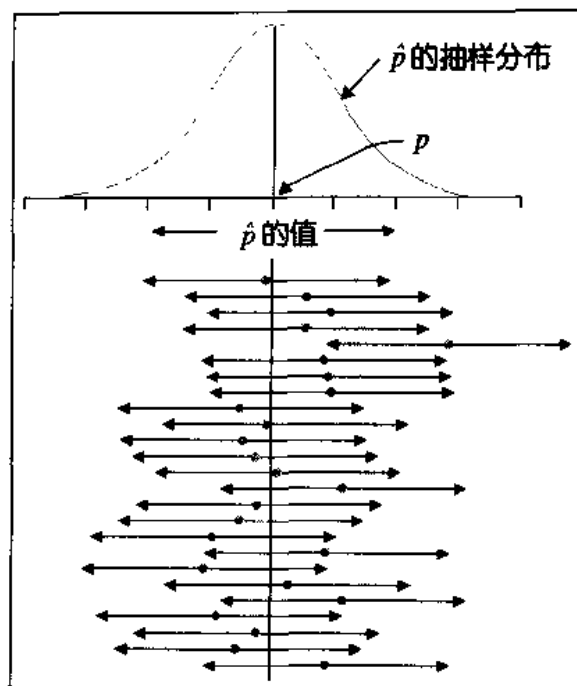


图 21.4 从同一总体抽出的 25 个样本所产生的 95% 置信区间。长期来说，所有这样的区间中，有 95% 会包含真正的总体比例，在这个图里面由垂直直线来代表

还有，我们的方法有假设总体很大——至少要有样本大小的 10 倍大。专业的统计学家使用较复杂的方法，会把总体大小也纳入考虑，这些方法连小样本都可以适用。但是我们的方法在许多实际应用的情况下都已经够好了，更重要的是这个方法让我们学到，怎样可以从统计量的抽样分布找到置信区间。而任何置信区间背后的道理都是这样的。

总体比例的置信区间*

我们用了 68-95-99.7 规则的 95 部分，得到总体比例的 95% 置信区间。也许你觉得一个只有在 95% 的时候管用的方法还不够好，你希望能有 99% 的信心。这样必须先找到正态分布的中间 99% 是在哪里。对任意在 0—1 之间的概率 C ，都存在一个数 z^* ，使得任

* 此节为选读。

何正态分布在平均数两侧 z^* 个标准差范围内的概率是 C 。图 21.5 里表示出概率 C 和 z^* 这个数之间的关系。

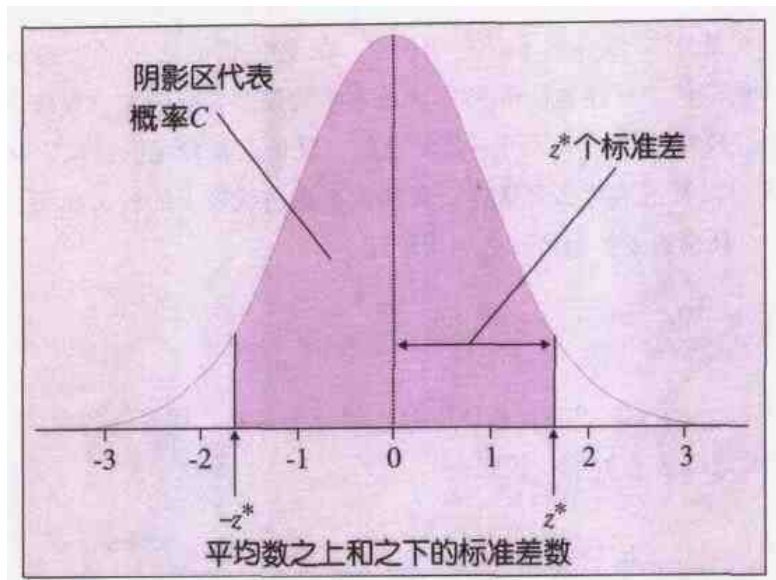


图 21.5 正态分布的临界值 z^* 。在任何正态分布中，平均数在 $-z^*$ 和 z^* 个标准差范围内，曲线之下的面积(概率)是 C

表 21.1 里有不同 C 值对应的 z^* 值。为了方便应用，表里面的 C 是用置信水平的百分比表示。 z^* 这个数叫做正态分布的临界值 (critical value)。从表 21.1 里可以看得出，任意正态分布在平均数 ± 2.58 个标准差范围内的概率是 99%。也可看出任意正态分布在平均数 ± 1.96 个标准差范围内的概率是 95%。68-95-99.7 规则中，用 2 来代替了临界值 $z^* = 1.96$ ，就实际应用来说这样够好了，但表里面的是更精确的值。

表 21.1 正态分布的临界值

置信水平 $C(\%)$	临界值 z^*	置信水平 $C(\%)$	临界值 z^*
50	0.67	90	1.64
60	0.84	95	1.96
70	1.04	99	2.58
80	1.28	99.9	3.29

从图 21.5 里可以看出，样本比例 \hat{p} 的值会落在 p 的 z^* 个标准差范围内的概率是 C 。这也就是说，从观测到的 \hat{p} 值往两侧各延伸 z^* 个



标准差所得到的区间，会抓到未知的 p 的概率是 C 。 \hat{p} 的标准差用估计值代替，就可以得到以下公式。

比较例 5 和例 3 中 95% 置信区间的计算过程，会发现惟一的差别，就是 95% 信心水平时所用的 2，在 99% 信心水平时被临界值 2.58 所取代。这样做使得 99% 置信水平的误差界限较大，置信区间较宽。较高的信心水平可不是免费的，代价就是较宽的区间。从图 21.5 可以看出为什么会这样。要涵盖正态曲线底下的较大面积，从中心点往两边走的距离就必须比较远。

• 总体比例的置信区间

从一个总体中抽取一个大小为 n 的 SRS，总体中有比例 p 为成功。样本中的成功比例为 \hat{p} 。当 n 够大时， p 的近似水平 C 置信区间为

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

式中的 z^* 是表 21.1 中对应概率 C 的临界值。

例 5 99% 置信区间

“全国艾滋行为调查”的 2 673 位成年同性恋者的随机样本中，有 170 人在前一年有不止一个性伴侣。我们想要找所有同性恋者中有多个性伴侣的比例 p 的 99% 置信区间。从表 21.1 知道，对应 99% 置信水平，我们必须延伸 $z^* = 2.58$ 个标准差。以下是计算过程：

$$\begin{aligned}\hat{p} &= \frac{170}{2\,673} = 0.063\,6 \\ \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= 0.0636 \pm 2.58 \sqrt{\frac{(0.063\,6)(0.936\,4)}{2\,673}} \\ &= 0.0636 \pm (2.58)(0.004\,7) \\ &= 0.0636 \pm 0.012\,1 \\ &= 0.051\,5 - 0.075\,7\end{aligned}$$

我们有 99% 信心，真正的总体比例在 5.15%—7.57% 之间。也就是说，我们得到这个百分比范围所用的方法，有 99% 的时候会产生正确结果。



本章重点摘要

统计推论根据样本里的数据，对总体做结论。因为我们没有整个总体的数据，所以结论并不是完全确定的。**置信区间**估计一个未知参数的方式，可以提供我们该估计的不确定程度。区间本身就已告诉我们，对未知参数可以“定位”到什么程度。**置信水平**是一项概率，它告诉我们在许多样本当中，我们的方法所产生的区间确实会抓到参数的机会有多大。要找置信区间，先得考虑统计量的**抽样分布**，也就是重复抽样之下统计量会如何变化。

在本章中我们讨论了特定的一种置信区间，也就是根据从总体抽出的 **SRS**，得到总体中“成功”比例 p 的置信区间。第 23 章中还有对解释置信区间的更多建议。



第21章 习题

21.1 学生意见调查。唐雅想要估计她宿舍里的同学，有多少比例喜欢宿舍提供的食物。她在住宿舍的 175 位学生中，访问了 50 人的 SRS。其中有 14 个人认为宿舍供应的食物不错。

- (a) 唐雅想要对怎样的总体做结论？
- (b) 用你自己的话来说，这题里的总体比例 p 是指什么？
- (c) 从唐雅的样本得到的样本比例 \hat{p} 的值是多少？

21.2 该开除教练吗？一所专科学校的校长说：“校友中有 99% 支持开除柏格斯教练的决定。”你联系了该校尚在人间 的 15 000 名校友中 200 人的 SRS，并且得知只有 76 人赞成开除教练。

- (a) 这里要做的推论是关于什么总体？
- (b) 仔细说明这题里的总体比例 p 指的是什么？
- (c) 样本比例 \hat{p} 的值是多少？

21.3 学校最严重的问题。一项访问了 1 500 名成人的抽样调查报告说：“在 95% 置信水平之下，所有美国成人当中，有 27%–33% 的人认为，全国公立学校面对的最严重问题是毒品。”说明报告中的“95% 置信水平”是什么意思，请假设你的对象不懂统计。

21.4 枪支暴力。哈里斯调查 (Harris Poll) 问了 1 009 人的成人样本，哪些死因他们认为在将来会更为普遍。枪支暴力夺得冠军，样本中有 706 人认为受枪击致死的事件会增加。虽然全国调查用的样本不是 SRS，但是够接近 SRS 了，因此用我们的方法还是可以得到大致正确的置信区间。

- (a) 本调查的总体比例 p 是什么，用话说出来。
- (b) 算出 p 的一个 95% 置信区间。
- (c) 哈里斯宣布此次调查结果的误差界限是正负 3 个百分点。你在 (b) 中做出的结果和这个误差界限差别大不大？

21.5 对女性抽样。《纽约时报》一项对女性议题的调查，访问了从全美国随机抽样的 1 025 位女性，但阿拉斯加和夏威夷除外。样本



中的女性，有 482 位说她们没有足够的个人时间。虽然全美国调查的样本不是 SRS，但是够接近了，用我们的方法还是可以得到大致正确的置信区间。

- (a) 本题中的参数 p 是什么，请用文字说明。
- (b) 利用调查结果，找出 p 的 95% 置信区间。
- (c) 对 (b) 中做出的结果做简短说明，假设你要解说的对象不懂统计。

21.6 枪支暴力。在习题 21.4 中，你根据一个 $n = 1\ 009$ 位成人的随机样本算出了 95% 置信区间。如果希望误差界限只有习题 21.4 中的一半大，需要用多大的样本？

21.7 样本大小的影响。一项民意调查发现，样本中有 60% 的人认为平衡联邦预算比减税重要。算出所有成年人中有同样想法者所占比例的 95% 置信区间，假设结果 $\hat{p} = 0.6$ 来自以下大小的样本：

- (a) $n = 750$
- (b) $n = 1\ 500$
- (c) $n = 3\ 000$
- (d) 从你的结果可看出增加样本大小有何影响？请简短说明。

21.8 随机数字。我们知道在一大群的随机数字当中，0 所占的比例是 $p = 0.1$ ，因为所有 10 个可能数字的机会都相同。随机数字表中的数字，是所有随机数字总体中的一个样本。要抽取 200 个随机数字 SRS，可以在书末的表 A 中的列 101—列 125，每 5 个数字一组总共 200 组，取每组第一个数字。这 200 个数字中有几个 0？算出这个数字样本所来自的总体中，0 所占比例的 95% 置信区间。你的区间有没有涵盖真正的参数值，也就是有没有包含 $p = 0.1$ ？

21.9 掷图钉。如果你把一颗图钉扔在硬的平面上，它“着陆”时尖头会朝上的概率是多少？拿一颗图钉来扔 100 次，估计这个概率。你的这 100 次投掷，是来自所有可能投掷的总体，大小为 100 的 SRS。100 次投掷，尖头朝上的比例，就是样本比例 \hat{p} 。用你掷的结果来造一个 p 的 95% 置信区间。针对你做出的结果写一段简短说明，你所要解说的对象是不懂统计，却想知道图钉到底有多大的概率会尖端朝上的人。



21.10 别忘了基本要点。全国艾滋行为调查的结果是，在 2 673 位成年异性恋者的随机样本中，有 170 人承认在前一年当中有多个性伴侣。我们利用这个结果，算出了所有成年异性恋者中，有多个性伴侣者所占比例的置信区间。这项抽样调查可能隐含一些偏差，是我们的置信区间没有列入考虑的，为什么可能有偏差？样本比例 6.36% 大概是高估还是低估了真正的总体比例？

21.11 布方伯爵的铜板。18 世纪的法国自然主义者布方伯爵曾掷了 4 040 次铜得板，到 2 048 个正面。替布方伯爵的铜板正面朝上的概率造一个 95% 置信区间。你会不会很有把握这个概率不是 $1/2$ ？为什么？

21.12 十几岁青少年的电视机。《纽约时报》和 CBS 新闻主导了一项全国调查，访问了随机选出的 1 048 位 13—17 岁的青少年。这些青少年当中，有 692 人在自己房间里有电视机，189 人选福克斯 (Fox) 为他们最爱的电视频道。我们可以把样本当作 SRS。

- (a) 替在调查的那段期间，所有这个年龄层的人，在自己房间有电视机者所占比例，以及最爱福克斯频道的比例，分别造 95% 置信区间。
- (b) 新闻报道说明：“理论上来说，20 次当中有 19 次，调查结果和所有美国 13—17 岁青少年真正结果的差距，不论正负都不会超过 3 个百分点。”说明你的结果是否符合这段叙述。

21.13 哈雷摩托。哈雷摩托车在全美领有牌照的摩托车当中，占了 14%。你计划要访问 600 位摩托车车主的 SRS。

- (a) 你的样本中拥有哈雷的比例的抽样分布是什么？
- (b) 你的样本中，拥有哈雷者占至少 18.2% 的机会会有多大？样本至少包含 11.2% 哈雷拥有者的机会又有多大？利用 68-95-99.7 规则和你对 (a) 的答案来回答问题。

21.14 你慢跑吗？假设成年人中有 15% 常慢跑。一项意见调查询问了 500 位成人的 SRS，问他们是否慢跑。

- (a) 样本中的慢跑者比例 \hat{p} 的抽样分布是什么？
- (b) 根据 68-95-99.7 规则，样本中的慢跑者至少占 11.8% 比例的概率是多少？



21.15 速算法。第 3 章的速算法用 $\hat{p} \pm 1/\sqrt{n}$ 当作总体比例 95% 置信区间的粗算公式。速算法算出的误差界限，比实际上大一些。它和本章的较精确方法，在 \hat{p} 靠近 0 或 1 时，差别会最大。在登记有案的摩托车中抽取的 500 辆的 SRS，其中有 68 辆为哈雷摩托车。找出所有摩托车中哈雷所占比例的 95% 置信区间，用速算法和本章方法各算一次。速算法的误差界限会大多少？

21.16 68% 置信水平。我们利用 68-95-99.7 规则的 95 部分，来导出总体比例 p 的 95% 置信区间的公式。

- (a) 利用该规则的 68 部分，来找出 68% 置信区间的公式。
- (b) 用简单易懂的语言解释“68% 置信水平”是什么意思。
- (c) 用全国艾滋行为调查的结果(例 3)，算出有多个性伴侣的异性恋者所占比例的 68% 置信区间。你的区间和例 3 中的 95% 区间相比，有何差别？

21.17 模拟置信区间。在习题 21.16 里，你导出了总体比例 p 的 68% 置信区间的公式。假设美国国会议员柯可丝的选区中，有 60% 的选民支持她竞选连任(这个百分比没有人知道)。

- (a) 你会怎样模拟 25 个选民的 SRS 的投票情况？
- (b) 模拟 10 个 SRS*，每个 SRS 用表 A 里不同的列。你得出来支持柯可丝的样本比例 \hat{p} 的 10 个值是多少？
- (c) 根据你的 10 个样本，分别找出 p 的 68% 置信区间。10 个区间当中有几个抓到了真正的参数值 $p=0.6$ ？(大小为 25 的样本不够大，不能期望结果很准，但是即使是小规模模拟，也可以看清楚在重复抽样时，置信区间会有什么变化。)

* 译注：即 10 回合，一回合代表 25 位选民的投票情况。

以下习题对应本章中选读部分的内容。

21.18 枪支暴力。习题 21.4 中报告了一项哈里斯调查的结果；在 1 009 位成人的随机样本中，有 706 人认为枪击致死事件在未来会增加。替所有成人中有这样想法的比例造一个 90% 置信区间。你的区间和习题 21.4 中的 95% 置信区间比起来，有何差别？

21.19 对女性抽样。习题 21.5 中报告了一项《纽约时报》的民意调查结果：在 1 025 位女性的随机样本中，有 482 人觉得没有足够的



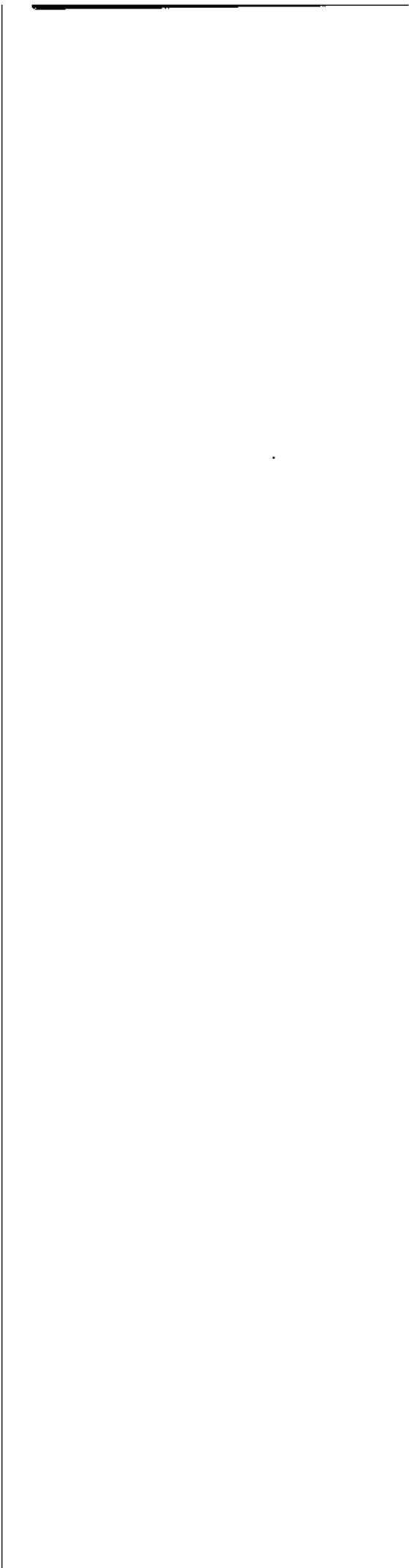
个人时间。替所有女性中有这种想法者所占的比例，造一个 99% 置信区间。你的区间和习题 21.5 中的 95% 置信区间比起来，有何差别？

21.20 置信水平的影响。一个 1 500 位美国成人的随机样本中，有 60% 认为平衡联邦预算比减税重要。利用此项调查结果以及表 21.1，替所有成人中有同感者的比例，造出 70%、80%、90% 及 99% 的置信区间。从你的结果看起来，改变置信水平会发生什么影响？

21.21 不满意的 HMO 病人。曾向卫生维护组织 (HMO, health maintenance organization) 提出不满申诉的病人，退出该组织的机会有多大？最近有一年当中，新英格兰地区的一家大型保健组织，在超过 400 000 名的会员中，有 639 人提出不满申诉，而其中有 54 人自动退出了 HMO (也就是说并不是因为搬家或换工作而被迫退出。) 把这一年的申诉者当作所有以后会申诉的病人中的一个 SRS。造一个申诉会自动退出 HMO 的比例的 90% 置信区间。

21.22 估计失业率。美国劳工统计局用 90% 置信区间来呈现每个月当前人口调查 (CPS) 的失业数据。2000 年 1 月的调查，访问了群众中属于劳动人口的 133 357 人，其中有 6 264 人没有工作。CPS 的样本并不是 SRS，但是为了做这习题，就当作劳工统计局抽了 133 357 人的 SRS。算出失业率的 90% 置信区间 (失业率指所有属于劳动人口的人中，没有工作者所占比例)。

21.23 安全的误差界限。误差界限 $z^* \sqrt{\hat{p}(1-\hat{p})/n}$ 在 \hat{p} 等于 0 或 1 时是 0，而在 \hat{p} 为 1/2 时的值最大。想要了解为何如此，可以计算一下当 \hat{p} 等于 0、0.1、0.2……0.9 及 1 时， $\hat{p}(1-\hat{p})$ 的值。把 \hat{p} 的值当作横座标， $\hat{p}(1-\hat{p})$ 的值当作纵坐标，画出这些点，再画一条曲线通过这些点。你已经画出了 $\hat{p}(1-\hat{p})$ 的图。这个图是不是在 $\hat{p}=1/2$ 时达到最高点？这也就是说，用 $\hat{p}=1/2$ 所算出来的误差界限，必定至少和实际的误差界限一样大。



第 22 章

什么是显著性检验

这个社区完了？

即便有公平居住条例及其他法律条例，美国大部分黑人及白人仍然隔离而居。当一个社区的黑人比例过高时，白人常常会搬走。多年以来白人的态度是否有改变？现在，当黑人家庭陆续迁入时，是不是有更多白人愿意留下？

底特律区域研究 (Detroit Area Study) 于 1976 年访问了底特律大都会区 1 104 位成人的随机样本，又在 1992 年另外访问了 1 543 位成人的随机样本。底特律是美国大城市中隔离问题最严重的，所以该项研究仔细探讨了对于不同种族住在同一社区的看法。有一个问题是要白种人先想像他们住在一个全是白人的区域，像图 22.1 的第一张卡

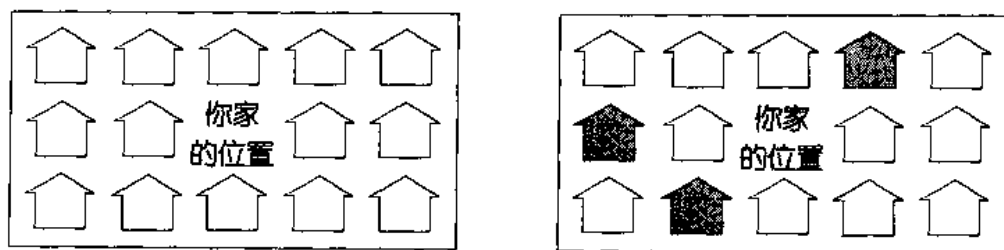
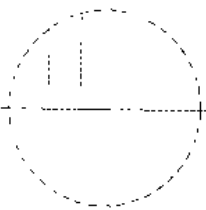


图 22.1 底特律区域研究给白人居民看过两张卡。他们要求这些白人先想像他们居住的区域是像左边这样。然后有些黑人搬来，住家附近变成像右边这种情况。然后询问这个白人家庭会不会想要搬走

所显示的。对这项研究有回应的人中，的确大部分都是住在全白人社区。然后访问者给他们看图 22.1 的第二张卡，卡上的 15 栋房屋中有 3 栋是由黑人居住，这是整个底特律地区真正的黑人比例。他们会想搬离这样的社区吗？

1976 年的样本中，有 24% 的白人会想搬走。到 1992 年，这个百分比降到了 15%。看起来似乎态度上的改变使得“白人迁徙”活动较少发生。（这里假设回应者说的是实话。也有可能是从 1976 年到 1992 这段期间，大家对赞成隔离的言论，接受度变低了。）这些结果是根据两个适度大小的样本得来的。这两个样本的差别，可不可能只是因为随机选择受访者而碰巧发生的？提出该研究报告的人回应了这项挑战，提出了两个随机样本完全因为机遇，产生像 24% 对应 15% 这么大差距的概率，这个概率很小，还不到 0.01。所以观察到的差距“有统计显著性”（statistically significant）。这个概念，以及用像 0.01 这样的概率来代表这个概念，就是我们在本章要讨论的。



统计检验的理论基础

常常在附近打休闲篮球的一个自以为是的球员号称，他的罚球命中率有 80%。你对他说：“投给我看看。”他投了 20 球，结果投中 8 球。“啊哈！”你下了结论：“如果他的命中率真是 80%，那他几



乎不大可能会在投 20 次时只中 8 球。所以我不相信他的话。”这就是统计检验(statistical test)背后论理基础的休闲版本；在断言正确时很少会发生的结果若发生了，就是断言不正确的证据。

统计推论是利用样本的数据，来对总体做结论。所以正式来说，统计检验处理的是有关总体的断言。检验要判断的是，样本数据是否提供了不利于断言的证据。检验说的是：“如果我们取许多样本而且断言正确，我们很少会得到这样的结果。”要得到样本证据强度的数值量度，就要把语意模糊的“很少”用概率来取代。我们从下面的例子来看看，该怎么运用这种推理过程。

例 1 咖啡是现煮的吗？

注重口味的人，想来应该是喜欢现煮咖啡超过即溶咖啡的。但从另一方面来看，有些喝咖啡的人也可能只是对咖啡因有瘾。一位持怀疑态度的人断言：喝咖啡的人里，只有一半偏好现煮咖啡。让我们做个实验来检定这个断言。

让全部 50 个受试对象都品尝两杯没做记号的咖啡，并且要说出喜欢哪一杯。两杯中有一杯是即溶咖啡，另一杯是现煮咖啡。由实验结果得到的统计量，算的是样本中说比较喜欢现煮咖啡的人的比例 \hat{p} 。我们发觉：50 位受试对象中，有 36 位选的是现煮咖啡。也就是：

$$\hat{p} = \frac{36}{50} = 0.72 = 72\%$$

为了清楚说明，让我们把结果 $\hat{p} = 0.72$ 和另一个可能结果做比较。如果 50 位受试对象里，只有 28 位喜欢现煮咖啡超过即溶咖啡，样本比例就是

$$\hat{p} = \frac{28}{50} = 0.56 = 56\%$$

当然用 72% 这个数据来否定这位怀疑者的断言，比起用 56% 要强。但是强多少呢？即使样本里有 72% 的人喜欢现煮咖啡，但这就可以当做总体中大部分人都是如此的有力证据吗？统计检验可以回答这些问题。

下面是用概要形式来回答问题：

- **断言。**怀疑的人断言：喝咖啡的人里，只有一半偏好现煮咖啡。换句话说，他断言总体比例 p 只有 0.5。为了方便讨论，假设这个断言是正确的。



- **抽样分布。**如果断言 $p=0.5$ 是正确的，而我们检验了很多个包含 50 位喝咖啡的人的随机样本，样本比例 \hat{p} 的值会随样本而变化，遵循(近似)正态分布，其

$$\text{平均数} = p = 0.5$$

及

$$\begin{aligned}\text{标准差} &= \sqrt{\frac{p(1-p)}{n}} \\ &= \sqrt{\frac{(0.5)(0.5)}{50}} \\ &\approx 0.0707\end{aligned}$$

图 22.2 里画出了这条正态曲线。

- **数据。**把样本比例 \hat{p} 的值标示在抽样分布上。你可以从图 22.2 看出来， $p=0.56$ 这个值很正常，但是 $p=0.72$ 就很稀奇了。如果喝咖啡的人里只有 50% 喜欢现煮咖啡，则在 50 位喝咖啡者的样本中，出现 72% 的人喜欢现煮咖啡的情况会非常少见。所以样本数据的确提供了不利断言的证据。
- **概率。**我们可以用概率来度量，对断言不利的证据到底有多强。当总体的真正比例是 $p=0.5$ 时，一个样本的 \hat{p} 值会这么大或更大的概率是多少？若 $\hat{p}=0.56$ ，这个概率就是图 22.2 中正态曲线之下的阴影区面积。这个面积是 0.20。我们的样本比例值事实上是 $\hat{p}=0.72$ ，而只有 0.001 的几率会得到这样大的样本结果，它对应的区域小到在图 22.2 里根本看不到。在所有样本中，光因为机遇就有 20% 会发生的结果，无法当成断言不正确的有力证据。但是在 1 000 次当中只会发生一次的结果，就是很好的证据。

要确定你真的了解为什么这个证据令人信服。有两种可能的解释，可以说明为什么会得到“受试对象中有 72% 比较喜欢现煮咖啡”的这个结果：

- (a) 怀疑者是对的($p=0.5$)，但是因为运气太差，应该极不可能发生的结果却发生了。
- (b) 事实上偏好现煮咖啡的总体比例大于 0.5，所以样本结果差不多就是预期的结果。

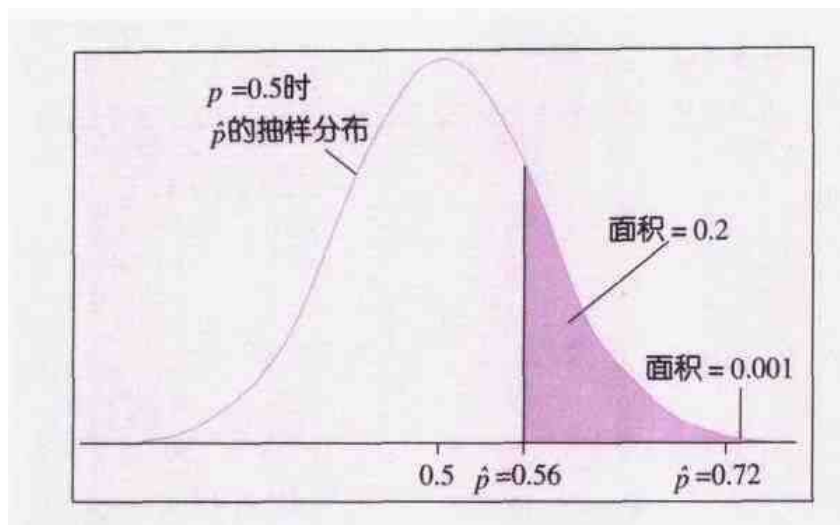


图 22.2 50 位喝咖啡的人当中，喜欢现煮咖啡者所占比例的抽样分布。这个分布成立的前提是，所有喝咖啡者中有 50% 喜欢现煮咖啡。阴影区的面积即是样本比例至少是 56% 的概率

我们不能确定(a)一定不对，因为我们的口味测试结果有可能真的就只是机遇造成的。但是，这样的结果完全是由机遇造成的概率非常小(0.001)，所以我们相当有信心的认为(b)才是对的。

假设及 P 值

显著性检验对这个基本论据，处理得更精确(或者说隐藏起来)。在大部分的研究中，我们想要证明：总体中有某种特定的效应。例如在例 1 中，我们猜想大部分喝咖啡的人偏好现煮咖啡。为方便讨论，统计检验会先假设我们在找的效应并不存在。然后我们开始寻找的证据，必须不利这个假设，而支持我们想找的效应。显著性检验的第一步，是要先列出一个断言，然后我们再试着找出证据来否定它：

• 原假设 H_0

在统计检定中，受检验的断言叫做**原假设**(null hypothesis)。检验是设计来评估，否定原假设的证据有多强。通常，原假设都是“没有效应”或“没有差别”的叙述。



“原假设”这个词用 H_0 代表，读法是“H 零”。 H_0 是关于总体的叙述，所以一定要用总体参数来表示。例 1 当中的参数是所有喝咖啡的人里面，偏爱现煮咖啡的比例 p 。原假设是：

$$H_0: p = 0.5。$$

我们希望或猜想可以取代 H_0 的正确叙述，叫做备择假设 (alternative hypothesis)，用 H_a 表示。在例 1 中，备择假设就是喝咖啡的人，大多偏好现煮咖啡。用总体参数表示就是：

$$H_a: p \neq 0.5。$$

显著性检验会找对原假设不利，但对备择假设有利的证据。如果观测到的结果，在原假设为真的情况下是出人意料的，而在备择假设为真时却较易发生，这个证据就很强。比如说，当事实上总体只有一半喜欢现煮咖啡时，发现 50 位受试对象中有 36 位喜欢，就会出人意料。有多么出人意料呢？显著性检验用概率来回答这个问题：这个概率指的就是，在 H_0 正确时得到的结果跟预期结果的差距很大的概率，而这个差距至少要等于（或大于）实际观测与预期结果的差距。怎么样算是“跟预期结果的差距很大”？这既和 H_0 有关，也和 H_a 有关。在口味测试中，我们希望得到的概率，就是在 50 人中至少有 36 人喜欢现煮咖啡的概率。如果原假设 $p = 0.5$ 正确的话，上述的这个概率会非常小 (0.001)。这就是原假设不正确的有力证据。

• P 值

统计检验的 **P 值 (P-value)** 是在 H_0 为真的假设下，得到样本结果会像实际观测结果那么极端或更极端的概率。 P 值愈小，资料所提供否定 H_0 的证据就愈强。

在实际应用时，大部分的统计检验可以由会计算 P 值的电脑软件来执行。在许多领域中都常见到引用 P 值来描述研究结果。所以，即使你自己不做统计检验，也应该要知道 P 值的意义；就像虽然你自己不用计算置信区间，也应该要了解“95% 信心”是什么意思。

逮到你了！

查税员怀疑“剥削公司”常常开假支票来给支出灌水，以达到减税的目的。她想查出真相但不想检验每一张支票，于是用电脑帮忙。真实数据的第一个数字遵循着人所共知的形态，而并不是从 0 到 9 每个数字概率都一样。如果支票上的金额不符合这个形态，她就会深入调查。同一条街上，有黑客正在探查某家公司的电脑档案，因为档案经过加密，所以他没法读取。但他还是找得到解码的钥匙——就是惟一每个符号概率都一样的长串符号。



布方伯爵铜板投掷实验的真相

例 2 布方伯爵的铜板

法国自然主义者布方伯爵掷了 4 040 次铜板，他得到 2 048 次正面。正面的样本比例是：

$$\hat{p} = \frac{2\,048}{4\,040} = 0.507$$

这个结果比一半稍多一点。这是不是布方的铜板不平衡的证据呢？这就是显著性检验可以发挥作用的时候了。

假设。原假设说铜板是平衡的 ($p=0.5$)。在我们看到资料之前，并没有怀疑铜板会偏向哪个特定方向，所以备择假设只是“铜板不平衡”。两个假设分别是

$$H_0 : p = 0.5$$

$$H_a : p \neq 0.5$$

抽样分布。如果原假设为真，样本中的正面比例就会近似正态分布，其

$$\text{平均数} = p = 0.5$$

$$\begin{aligned}\text{标准差} &= \sqrt{\frac{p(1-p)}{n}} \\ &= \sqrt{\frac{(0.5)(0.5)}{4\,040}} = 0.007\,87\end{aligned}$$

图 22.3 画出了这个抽样分布，并标示出布方的样本结果 $\hat{p}=0.507$ 。光看图就看得出，样本结果很正常，无法提供证据来否定 $p=0.5$ 的断言。

P 值。得到的结果和 0.5 的差距，会至少有布方的 $\hat{p}=0.507$ 远的机会有多大？因为备择假设里 \hat{p} 可能在 0.5 之左，也可能在 0.5 之右， \hat{p} 值往左、右任一方向远离 0.5，都提供了否定 H_0 而肯定 H_a 的证据。因此，P 值是观测值 \hat{p} 在左、右任一方向，偏离 0.5 的程度至少和 $\hat{p}=0.507$ 相同的概率。图 22.4 用正态曲线底下的面积来表示这个概率。它是 $p=0.37$ 。

结论。在布方试验的不断重复当中，真正平衡的铜板会有 37% 的时候，得到离 0.5 这么远或更远的结果。布方的结果让我们没有理由认为他的铜板不平衡。

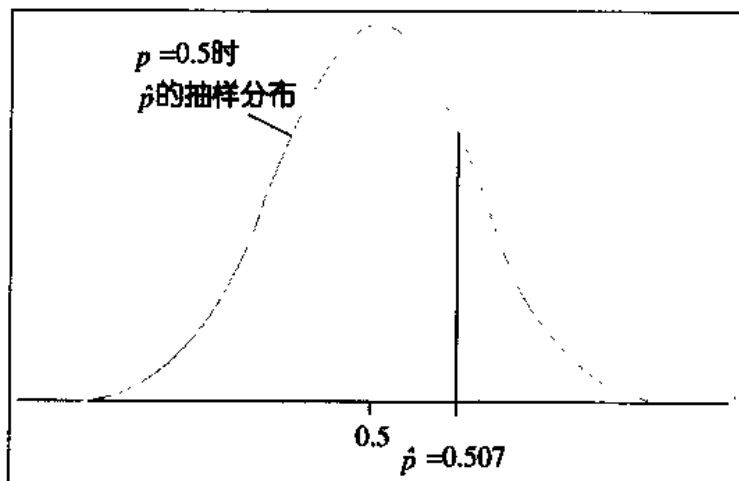


图 22.3 掷平衡铜板 4 040 次，正面比例的抽样分布。布方伯爵得到的 0.507 的正面比例结果，标示如图

例 1 中的备择假设 $H_a: p > 0.5$ 是单边备择假设 (one-sided alternative)，因为我们寻找证据是期望能够说：总体比例大于 1/2。例 2 中的备择假设 $H_a: p \neq 0.5$ 是双边备择假设 (two-sided alternative)，因为



我们只问铜板是否平衡。

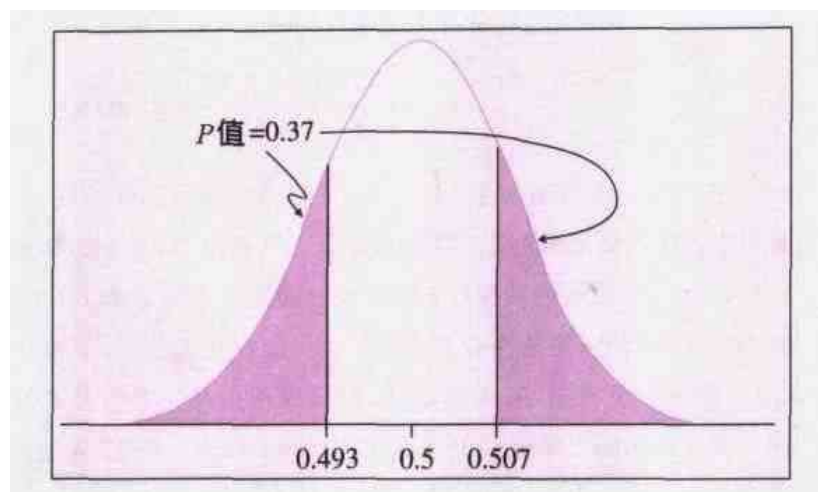


图 22-4 检定布方伯爵的铜板是否平衡所得到的 P 值。这是假设铜板平衡时，样本比例距 0.5 的距离，会至少像布方的结果 0.507 这么远的概率

因此，样本结果到底要往单向还是双向偏离，才可以算做否定 H_0 而肯定 H_a 的证据，要根据备择假设是单边还是双边来决定。

统计显著性

我可以在事前决定，用于否定 H_0 的证据必须强到何种程度。这等于是说我们要求多小的 P 值，而这个关键的 P 值就叫做显著水平 (significance level)，通常用希腊字母 α (读做 alpha) 表示。假如我们选择 $\alpha = 0.05$ ，我们要求的是：资料所传达否定 H_0 的证据要强到，当 H_0 正确时这种结果发生的频率不超过 5% (20 次中发生 1 次)。如果我们选 $\alpha = 0.01$ ，我们就是坚持要有否定 H_0 的更强证据，证据要强到，当 H_0 事实上为真时，这种结果只有 1% 的时候会发生 (100 次中有 1 次)。

统计显著性

如果 P 值小于或等于 α 值，我们称该组数据于有 α 的统计显著性水平 (statistically significant at level α)。



“显著”(significant)在统计上的意义并不是“重要”，而只代表“光是靠机遇不容易发生”。我们在第 5 章中用过这个字眼。现在我们给统计显著性加上数字，来表示所谓“不容易”到底是什么意思。有 0.01 的显著水平，常常是用下列叙述表示：“结果有显著性($P < 0.01$)。”这里的 P 代表 P 值。

我们并不需要用传统的 5% 和 1% 这些显著水平。 P 值提供了更多信息，因为 P 值让我们可以对我们选择的任意水平评估结果是否有统计显著性。举例来说， P 值为 0.03 的结果有 $\alpha = 0.05$ 的显著水平，但是没有 $\alpha = 0.01$ 的显著水平。然而传统的显著水平已经广为接受，做为“多少证据才足够”的标准。我们大概可以说： $P < 0.10$ 代表“有一些证据”不利原假设， $P < 0.05$ 代表“适度证据”，而 $P < 0.01$ 代表“有力证据”。不过可别太拘泥于这些准则。我在第 23 章还会谈到怎样解释检验结果。

计算 P 值*

要算出例 1 和例 2 中的 P 值，必须用到表 B 的正态分布百分位数来做正态分布的计算。这是第 13 章当中的选读部分。实际应用时可以用软件来计算，不过下面有个例子，示范怎样用表 B。

例 3 品尝咖啡

假设。例 1 中，我们想检验以下假设

$$H_0: p = 0.5$$

$$H_a: p > 0.5$$

此处的 p 是所有喝咖啡的人当中，喜欢现煮胜于即溶咖啡的人所占比例。

抽样分布。如果原假设为真，则 $p = 0.5$ ，而我们在例 1 中已见过， \hat{p} 遵循平均数为 0.5、标准差为 0.070 7 的正态分布。

数据。一个 50 人的样本中，有 36 人喜欢现煮咖啡。样本比例 $\hat{p} = 0.72$ 。

* 此节为选读。



P 值。备择假设是较大值那一头的单边假设。所以 P 值是会得到至少有 0.72 这么大的结果的概率。图 22.2 里将这个概率用正态抽样分布曲线之下的面积表示。要算任何正态曲线概率，先要把观测值标准化。结果 $\hat{p} = 0.72$ 的标准计分是：

$$\begin{aligned}\text{标准计分} &= \frac{\text{观测值} - \text{平均数}}{\text{标准差}} \\ &= \frac{0.72 - 0.5}{0.0707} = 3.1\end{aligned}$$

标准计分表 B 告诉我们，标准计分 3.1 是正态分布的 99.9 百分位数。这个意思是说，正态曲线之下，在 3.1 (标准计分) 左边的面积，是 0.999。因此右边的面积是 0.001，而这就是我们的 P 值。

结论。 P 值很小，表示数据提供了有力证据：大部分的人喜欢现煮咖啡。

网络寻奇

置信区间和统计检验在各种领域的研究报告中持续不断的出现。大部分的科学期刊，现在都至少把期刊中文章的摘要放到网站上。举例来说，第 21 章开头举的例子，出自医学期刊《循环》(circ. ahajournals.org)。本章开头的例子出自《美国社会学期刊》(*American Journal of Sociology*) (www. journals. uchicago. edu/AJS/home.html)。

在你的研究领域中选一本主要期刊。利用网络的搜索引擎找到该期刊的网站(只要搜寻期刊名称即可)。找几篇最近的论文看看内容摘要。如果你的领域所做的研究会产生数据，你大概一定会看到诸如“95% 信心”或“有显著性($P = 0.01$)”这类的句子。



本章重点摘要

置信区间估计一个未知参数。显著性检验则是要评估对某一未知参数断言的证据。实际执行时，统计检验的目的是要回答以下问题：

“我们从样本所看到的现象，只是因为机遇而碰巧发生的，还是总体中的确存在这种现象的有力证据？”

显著性检验用概率来回答这个问题，也就是光凭机遇就会得到像我们的样本这样极端的结果的概率有多少？这个概率就是 **P 值**。 P 值很小代表的意义是，我们的结果只靠机遇碰巧发生的机会不大。做检验之前，首先要写出**原假设**，这个假设表明，你所寻求的现象在总体中并不存在。**备择假设**陈述的是该现象存在。 P 值是在 H_0 为真的假设下所计算得到的，结果会向着备择假设的方向，达到像实际观测到的结果这么极端或者更极端的概率。如果一项样本结果在重复抽样情况下，光因机遇而发生的机会不超过 5%，则我们说这项结果有 **5% 的统计显著性水平**。

本章的内容是检验的基本论理基础，以及检验有关总体比例的假设的细节。在第 23 章中，对于如何解释统计检验的实际面有更多的讨论。



第22章 习题

22.1 种族优越感。一位社会心理学家报告说：“根据我们的样本，去教堂者的种族优越感，显著高于($P < 0.05$)不去教堂的人。”请对不懂统计的对象说明这是什么意思。

22.2 学生的收入 某大学负责奖助学金的单位向学生样本询问了他们的工作和收入状况。报告当中说：“以学年收入来说，不同性别间有显著差距($P = 0.038$)，男性平均来说收入较高。而黑人学生和白人学生之间却无差距存在($P = 0.476$)。”说明一下这两个结论，即性别对收入和种族对收入的影响，你用的语言要让不懂统计的人也听得懂。

22.3 饮食及糖尿病。多摄取纤维，能减低糖尿病患者的血中胆固醇含量吗？一项随机化临床试验比较了正常饮食和高纤饮食。以下是研究者的部分结论：“高纤饮食可使血中总胆固醇浓度降低%6.7($P = 0.02$)，三酸甘油脂的浓度降低%10.2($P = 0.02$)，及低密度脂蛋白胆固醇浓度降低%12.5($P = 0.01$)。”一位不懂统计的医师说，胆固醇减低6.7%不算什么，因为这也许是由于随机指派病人到这两种饮食，碰巧产生的结果。用一般的语言说为什么 $P = 0.02$ 可以反驳他这个说法。

22.4 饮食及肠癌。长久以来大家都认为，低脂高纤饮食可以减低罹患肠癌的风险。一项大型研究的结果却对这项建议提出了疑问。受试对象是2 079位曾在过去6个月内去除肠内息肉的人。这类息肉可能发展成癌。受试者随机分成两组，一组吃低脂高纤饮食，另一组是控制组，仍照平常的饮食习惯。接下来的4年中，息肉有没有复发呢？

(a) 描述这项实验设计的概要，可以用图表示。

(b) 令人意外的是，又长出息肉的状况“在两组之间没有显著差异”。清楚地说明这项发现是什么意思。

22.5 猪和声望在古中国的关系，在石器时代的中国，猪似乎不仅



仅是食物的来源而已。拥有猪也是富有的象征，这证据是从墓地的研究而来。如果祭祀用猪的头骨，常常伴随值钱的装饰品出现，就代表猪和装饰品一样，彰显着亡者的财富和声望。一项对公元前约 3500 年的埋葬研究做了以下结论：“在有猪头骨和没有猪头骨的墓地之间，陪葬物有非常大的不同……，一项检验显示，从所有工艺品中抽出两个样本，有显著水平为 0.01 的不同。”清楚说明为什么“有显著水平为 0.01 的不同”提供了我们充分理由相信，在有猪头骨和没有猪头骨的两块墓地之间，的确存在系统差异。

22.6 古埃及。要知道在法老王时代之前，埃及的定居地属于什么年代，是根据度量会随时间腐朽的某种碳的残余量来决定的。纳盖达 (Nagada) 区域的定居地第一次鉴定年代，是根据 60 年前挖掘出来的毛发。现在研究人员使用的是更新的方法，以及更晚出土的材料。鉴定出来的年代有差别吗？以下是关于某个位置的结论：“编号 KH6 的遗址有两种日期。从统计观点来看，两种日期没有显著差异。从这两种日期得出一个加权平均的修正年代为公元前 3715 ± 90 年。”向一位对古埃及感兴趣，但对统计没兴趣的人说明“没有显著差异”的意义。

22.7 礼物何价？人们对别人给他们的礼物，重视程度是否超过礼物的金钱价值？我们会盼望如此，因为我们希望“重要的是心意”。一项对 209 位成人做的调查，要求他们列出三项最近收到的礼物，并问他们：“除了感情上价值以外，在送礼的人永远不会知道的假设下，如果你可以得到的某个数目的金钱来取代该礼物，你最少要得到多少钱才会同样快乐？”结果大部分的人需要的金钱数目都超过礼物所值，才会一样快乐。魔术字眼“有统计显著性 ($P < 0.01$)”出现在这项发现的报告里。

- (a) 样本包含某研究所的学生及教职员，以及“在波士顿和费城的火车站及飞机场的一般社会大众”。报告说这个样本“并不理想”。你认为样本有多么不对？
- (b) 样本里的人认为礼物的价值“显著超过”礼物的实际价钱，代表什么意思？请用一般语言解释。
- (c) 现在要明确说明：“有统计显著性 ($P < 0.01$)”是什么意思？

22.8 上教堂。虽然一直以来，民意调查的结果都发现，有 40% 的



美国人说他们上周参加过宗教礼拜，但这几乎一定不可能是事实。

- (a) 为什么我们会预期调查结果会高估真正去教堂者的比例?
- (b) 你强烈地怀疑，任何一周中真正去教堂者所占比例其实低于40%。你计划抽一个成人的随机样本，再观察他们去不去教堂。你的原假设及备择假设各是什么?(一定要说清楚，你这项研究中的总体比例 p 是什么。)

22.9 体温。我们都听人说过，98.6°F(或37°C)是“正常体温”。事实上，有证据显示，大部分人的体温比这略低。你计划要抽一个随机样本，并精确度量里面每一个人的体温。你希望能够证明，大部分人的体温低于98.6°F。

- (a) 清楚说明这题当中的总体比例 p 代表什么。
- (b) 你的原假设和备择假设怎样用 p 表示?

22.10 失业率。上个月美国的全国失业率是4.3%。你觉得某个城市的失业率或许不同，所以计划做一项抽样调查，和“当前人口调查”问一样的问题。要判断当地的失业率是否和4.3%有显著差别，你会检验怎样的假设?

22.11 大学新生。你获知大一新生中有21%自认为是政治上的自由主义者。你觉得在你的学校里，这百分比不见得会一样，但是也不知会较高还是较低。你计划在你的学校对一年级生做抽样调查。要判断你的学校和全国结果是否有显著差别，你会检验什么假设?

22.12 我们的运动员有没有毕业?美国全国大学生运动员协会(NCAA)要求各高校报告各校运动员毕业的比例。在某个人型大学，1989—1991年之间入学的学生中，有70.7%在6年之内毕业。147位拿体育奖学金入学的学生中，有95位毕业。就把这147人当作所有在目前规则下会入学的运动员的一个样本。有没有证据显示，运动员毕业的比例不到70.7%?

- (a) 叙述说明，这题中的参数 p 是什么?
- (b) 原假设和备择假设 H_0 和 H_A 分别是什么?
- (c) 样本比例 \hat{p} 的值是多少? P 值是什么事件的概率?
- (d) P 值是 $P=0.053$ 。说明一下，为什么这代表有理由可以认为，



运动员的毕业比例比一般学生低。

22.13 我们想要有钱。最近有一年，在回应一项全国调查的大一新生中，有 73% 把“很有钱”当作重要的人生目标。某州立大学 200 位大一新生的 SRS 中，有 132 位认为这个目标很重要。我们想知道，是否该校大一新生中认为有钱很重要的比例，和全国比例 73% 有差别。

- (a) 叙述说明此题中的参数 p 是什么。
- (b) 原假设和备择假设 H_0 及 H_a 分别是什么？
- (c) 样本比例 \hat{p} 的值是多少？ P 值是什么事件的概率？
- (d) P 值是 $P=0.026$ 。详细说明，为什么这是合理足够的证据，指向 H_0 不对而 H_a 正确。

22.14 我们的运动员有没有毕业？习题 22.12 的结果，是否有 10% 的统计显著性水平？是否有 5% 的统计显著性水平呢？

22.15 我们想要有钱。习题 22.13 的结果是否是否有 5% 的统计显著性水平？是否有 1% 的统计显著性水平呢？

22.16 统计显著性水平有多大？用一般的用语来说明，为什么有 1% 的统计显著性水平的结果，必定也有 5% 的统计显著性水平。如果一项结果有 5% 的统计显著性水平，对于该结果是否有 1% 的统计显著性水平，你可以做什么结论？

22.17 统计显著性是指什么？一位学生在被问到“有 $\alpha=0.05$ 的统计显著性水平”，是什么意思时，回答说：“这个意思是说，原假设为真的概率小于 0.05。”这样解释是否正确？为什么？

22.18 用模拟来找 P 值。一个教小学一年级学生阅读的新方法(方法 B)，是否比目前使用的方法(方法 A)有效呢？你设计一项配对实验来回答这个问题。你把一年级学生配成了 20 对，每一对中的两个儿童在 IQ、社会经济地位及阅读准备测验分数各方面都很接近。你在每一对中随机指派一个儿童到方法 A，而另一个儿童就用方法 B 来教。一年级结束时，每个学生都要参加阅读能力测验。让 p 代表所有可能配对儿童中，用方法 B 教的儿童分数较高的比例。你的假设如下：



$H_0: p = 0.5$ (两个方法的效果没有差别)

$H_a: p > 0.5$ (方法 B 较有效)

你的实验结果是: 20 对里面有 12 对、方法 B 教的儿童分数较高, 即 $\hat{p} = 12/20 = 0.6$ 。

- (a) 如果 H_0 为真, 则 20 对学生可视为 20 个独立试验, 在每次试验中方法 B 赢的概率为 0.5。假设为了讨论方便, 我们假设 H_0 为真, 说明怎样可以用表 A 来模拟这 20 次试验。
- (b) 从表 A 的例 105 开始, 模拟此实验 10 个回合。用模拟结果来估计, 当 H_0 为真时, 方法 B 会在 20 对中的至少 12 对得胜的概率。(当然只模拟 10 个回合不足以对概率做出可靠的估计。不过只要你知道如何模拟, 多做几个回合是轻而易举的事。)
- (c) 说明为什么你在 (b) 中模拟的概率, 就是你的实验的 P 值。你要是有耐心的话, 可用类似这题的模拟方法, 找出这一节的所有 P 值。

22.19 用模拟来找 P 值 一项检测特异功能 (ESP, extra-sensory perception) 是否存在的标准实验, 是用一副有 5 种符号的牌来做的 (5 种符号分别是波浪、星星、圆形、方形及交叉)。当实验者把一张牌翻过去盖着, 并把注意力集中在牌上时, 受试者必须猜牌上的符号是哪一种。不具有 ESP 的受试者, 每次猜时单凭运气猜中的概率是 $1/5$ 。具有 ESP 的受试者, 猜中的比例就会比较高。茱莉在 10 次中说中了 5 次。(真正的实验会做很多次, 才能把弱 ESP 区别出来。在一长串的试验中, 从来还没有人对过一半这么多!)

- (a) 为了检验这个结果是否茱莉有 ESP 的显著证据, 先列出 H_0 及 H_a 。
- (b) 如果为了讨论缘故, 假设 H_0 为真, 说明怎样可模拟此实验。
- (c) 模拟此实验 20 回合, 从表 A 的列 121 开始。
- (d) 实际的实验结果是 10 次中有 5 次正确。此实验结果的 P 值, 是什么事件的概率? 根据你的模拟结果估计这个 P 值。茱莉的表现说服力有多强?

以下习题和计算 P 值有关, 该节内容属于选读部分。要执行一项检验, 必须如例 3 所示, 完成所有步骤 (假设、抽样分布、数据、图、计算)。

22.20 我们想要有钱。我们回头看看习题 22.13 的研究结果, 即



200 名大一新生中，有 132 人认为“很有钱”很重要，对这项数据的统计显著性，你能下怎样的结论？

22.21 青少年的电视机。《纽约时报》及 CBS 新闻主办了一项全国调查，访问了 1 048 位随机选出的 13—17 岁青少年。在这些青少年当中，有 692 人在自己房间里有电视机。我们可以把这样本当作 SRS。这项调查结果是否为合理证据，显示青少年中有超过一半的人在自己房间有电视机？

22.22 副作用 一项研究止痛药副作用的实验，将数种不需处方的止痛药分配给关节炎病人。使用了某一品牌止痛药的 440 位病人中，有 23 人产生了“不良副作用”。

(a) 如果所有病人中会有 10% 的人有不良副作用，则 440 名病人的样本中，有不良副作用者所占比例的抽样分布是什么？

(b) 服用此药之病人中，有不良副作用者所占比例不到 10%，此实验对于上列断言是否提供了有力证据？

22.23 化学家多生女儿？有些人相信，化学家生女儿的机会比一般人大。（有可能是因为化学家在实验室中暴露于某种化学品之下，影响了子女的性别。）美国华盛顿州的卫生局在出生证明上列出了父母的职业。在 1980—1990 年期间，有 555 名新生婴儿的父亲是化学家，而这些婴儿中有 273 名是女娃娃。同时期华盛顿州出生的所有婴儿中，有 48.8% 是女婴。有没有证据显示，化学家生女儿的比例，高于全州比例？

22.24 超速。我们常觉得路上开车的人似乎大部分都超速，当然视情况有所不同。不过这里有一组数据，记录研究者在马里兰州乡间一条州际公路上，观测到的驾驶员行为。该条公路的限速是每小时 55 英里，他们用一种埋在路下的电子装置记录车速，而且为了不计入大卡车，只考虑车身长度低于 20 英尺的车。他们发现 12 931 辆车中，有 5 690 辆超过速限。这是不是超速者不到一半的合量证据（至少在这个位置）？

第 23 章

统计推论的使用与滥用

马后炮专家

推论有其精妙之处，所以推论可能犯的错误也较细微。细微的错误不如明显的大错那么重要，所以不要忘记，大部分统计上的大错来源，和自发性回应样本、忽略潜在变量以及使用了无效量度等有关。不过推论还是会有错误或误导。而统计显著性更是如此。

我们来看看投资大众可以购买的 6 700 种共同基金。互联网中任何值得访问的投资网站，都会告诉你哪只基金在过去三年(或过去几年)内的获利最高。2000 年年初时，某网站声称获利最高的是动力网络基金。如果我在三年前买入这种基金的话，应该已经每年赚进了 112%。把这项获利和所有基金的平均获利相比，可看出明显高出许



多。

有件事应该很清楚，就是“回头找出表现最好的”并不适合做为显著性检验的根据。显著性检验的运作方式是先设好一项假设，诸如“动力网络的获利会超过平均”，然后再来等数据。若只是检视已有的数据，找出恰好在 6 700 只基金中表现最好的那只，再来问这支基金的获利是否超过平均，这问题一点意义也没有。

如果问题问的是，在一段日期表现优于平均的基金，是否会倾向于继续优于平均，这样问就有意义。如果答案是肯定的，去买绩优基金就有道理。金融教授在这个问题上花了大量的电脑工作时间，做了一大堆显著性检验。答案似乎是这样的：差基金通常会继续差下去（直到消失为止），但是却没有具统计显著性的证据，显示好的基金会一直好下去。事实上，动力网络基金在 2000 年的上半年中，位居 6 700 种共同基金中的最差 25 名当中。



聪明做推论

我们已见识过“置信区间”和“显著性检验”这两种主要的统计推论。不过我们在两种推论中都还只见过一种方法，也就是针对总体比例 p 所做的推论。你可以在许多书和软件当中，找到针对不同问题设置下，对各种参数做推论的各式各样方法。置信区间和显著性检验背后的理论基础还是一样的，只是细节可能看来很吓人。聪明做推论的第一步，是了解你的数据以及你想回答的问题，并斟酌适合问题架构的方法。以下是推论的一些要点，以我们熟悉的形式来呈现。

产生数据的设计很重要。“数据从哪儿来的？”仍是所有统计研究中应该问的第一个问题。任何推论方法都有特定的应用范围。以我们针对比例 p 所做的置信区间和检验来说：

- 数据必须是从我们关心的总体中所抽出的简单随机样本(SRS)。当你在用这些方法时，你其实是把数据当作 SRS 在处理。由于实际



退出实验

一项实验结果发现,要降低胆固醇和高血压,减轻体重显著比运动更为有效。170位受试者被随机指派到减肥计划、运动计划及控制组三者之一。170人当中只有111人完成整个计划,结论分析就只用了这111人的数据。退出者是否造成了偏差?在相信推论结果之前,一定要先弄清楚有关数据的各种细节。

上常常不可能真的从总体抽 SRS, 所以你的结论可能会受到挑战。

- 这些方法对于诸如分层样本这些比 SRS 复杂的抽样设计来说, 并不正确。对这些其他设计, 有别的方法可以处理。
- 对于随意搜集而得, 且偏差大小无法掌握的数据, 没有正确的推论方法。再好的方法也救不了差的数据。
- 其他误差来源比如中途退出者和无回应, 也都很重要。要记住置信区间和检验只会根据你喂进去的数据来得出结果, 对于上述的实际困难并不会列入考虑。

了解置信区间的运作。置信区间可以估计未知参数的值, 同时告诉我们估计的不确定程度有多大。所有的置信区间都符合以下的描述:

- 置信水平告诉我们的是, 同一个方法一再的使用, 其中会包含真正参数的比例。我们永远也不会知道, 我们手上这组数据所得到的区间, 究竟有没有包含真正的参数。我们能说的只是: “我得到这个结果, 是使用一种 95% 时候会包含真正参数的方法。”我们手上的数据, 有可能就属于算出的区间没包含真正参数的那 5%。如果你认为风险太大, 不妨改用 99% 置信区间。
- 高置信水平可不是平白得来的。根据同一组数据做的 99% 置信区间会比 95% 置信区间要宽。在估计参数的准确程度和对于包含参数的信心大小这两项之间, 只能尽量寻求平衡。
- 样本变大, 区间就会变窄。如果我们希望有高置信水平, 又要有较窄区间, 就必须取比较大的样本。 p 的置信区间的长度, 随着样本大小的平方根成比例的下降。要把区间的长度缩成一半, 观测数目就必须是原来的 4 倍。许多种类的置信区间大致都是这个情形。

了解统计显著性的意义。许多统计研究的目的, 是想要显示某种断言是正确的。临床试验将一种新药和标准用药比较, 因为医师希望新药对病人的帮助较大。研究性别差异的心理学家认为, 在一项度量建立人际关系网络的能力的测验当中, 女性的表现应该会比男性好(平均来说)。显著性检验的目的是要评估数据是否提供了足够证据, 可支持这类断言。也就是说, 检验帮助我们了解, 我们是否的确找到了正在寻找的东西。

要做到这点我们得知道, 若断言不正确的话会发生什么状况。这指的就是原假设: 两种药没差别, 女性和男性没差别。显著性检验只回答一个问题: “原假设不正确的证据有多强?” 检验



是用 P 值来回答这个问题。 P 值告诉我们，如果原假设正确的话，我们的数据会有多么不可能得到。相当不可能得到的数据，就是原假设不对的合理证据。我们永远也不会知道，对我们的总体来说这假设是否为真。我们能说的只是：“如果原假设为真，这样的数据只有 5% 的时候会发生。”

这类不利于原假设(但支持我们想找到的效应)的间接证据，不像置信区间那样直截了当。在下一节我还会多谈一些检验。

了解你用的方法必须满足的条件。我们对于比例 p 所做的检验以及置信区间，都要求总体必须比样本大很多。还要求样本本身也要够大，使得样本比例 p 的抽样分布会接近正态分布。我对于这些条件的细节说得不多，因为推论的论理基础更重要。就像有推论方法适用于分层样本一样，也有方法可以适用于小样本及较小总体。如果你要实际执行统计推论的话，需要找统计学家帮忙(或者必须多学很多统计学知识)，才能对付所有的细节。

我们当中的大部分人，读到统计研究结果的机会，要比自己处理数据的时候多。你要注意的是大问题，而不是作者是否用了百分之百正确方法的这类细节。这个研究是否问了正确的问题?数据从哪里来的?结果合不合理?研究结果中是否提出置信区间，让你不仅可以知道重要参数的估计值，还知道估计值的不确定程度?有没有提出 P 值来帮忙说服你，研究发现并不是碰巧得到的?

显著性检验面临的难处

显著性检验的目的，通常是想提出总体中存在某种效应的证据。这里说的效应，也许是指铜板的正面概率不是一半，或者用新的癌症疗法治疗的病人，有较长的平均存活时间。如果效应够大，就会在大部分样本中显示出来——我们掷铜板得到的正面比例会和一半相去甚远，或者用新疗法的病人会比控制组的病人多活很久。小的效应，比如正面概率和一半差不了多少，则通常会被样本的机遇变异给掩盖住了。这也理当如此：大的效应比较容易侦测出来。换句话说，当总体真正值离原假设很远的时候， P 值通常会很小。

检验的主要“弱点”，是它只度量不利于原假设的证据强度。检验并没有说我们正在寻求的总体效应，到底有多大或多重要。举例来说，我们的假设可能是“这个铜板是平衡的”。我们把这个假设，用



得到正面的概率 p 来表示成 $H_0: p = 1/2$ 。真正的铜板没有哪一个是百分之百平衡的，所以我们知道这项假设并不会完全正确。如果这个铜板的正面概率是 $p = 0.502$ ，从实际观点来看，我们可能认为它已经是平衡的了。可是统计检验可不管什么“实际观点”。它只会问是不是有足够证据显示， p 并不是恰好 0.5。检验把焦点放在不利于某个确切的原假设的证据强度上面，这点是应用检验时许多困扰的来源。

当你读一项显著性检验的结果时，要特别注意样本大小，理由如下：

- 较大的样本会让显著性检验比较敏感。如果我们掷铜板几十万次，则对于 $H_0: p = 0.502$ 的检验往往会得到很小的 p 值。检验结果并没有错（它找到合理的证据，显示 p 的确不是恰好 0.5。）但是它把这么小的差距也找出来，实在并没有什么实际用处。一项发现可能有统计显著性，却没有实际上的重要性。
- 另一方面来看，用小样本做的显著性检验，敏感度又常常不够。如果你掷铜板只掷 10 次，在检验 $H_0: p = 0.5$ 时，即使这个铜板真正的 p 是 $p = 0.7$ ，检验结果的 P 值也常常较大。这回检验仍然是正确的，因为只掷 10 次原本就不足以提供不利于原假设的合理证据。没有达到统计显著性不代表效应不存在，只能说我们没有找到合理证据来支持它。而小样本常常会漏掉总体中确实存在的效应。

不论总体的真实情况如何，不管是 $p = 0.7$ 或是 $p = 0.502$ ，观测值多一点，就可以让我们抓 p 的值抓得准些。若 p 不等于 0.5，观测值愈多就会给我们愈多证据，也就是较小的 P 值。因为显著性会受样本大小和总体的真正值的强烈影响，所以统计显著性并不能告诉我们，一项效应有多大或实际上有多重要。如果我们取的样本小，大的效应（比如当原假设为 $p = 0.5$ 时，实际上 $p = 0.7$ ），常常产生出未达到统计显著性的数据。如果我们取的样本很大，则小的效应（比如 $p = 0.502$ ）也常常产生出有高度统计显著性的结果。我们来看之前见过的一个例子，看看样本大小如何影响统计显著性。



例1 再论布方伯爵的铜板

布方伯爵掷一个铜板 4 040 次，得到 2 048 个正面。他的正面样本比例是

$$\hat{p} = \frac{2\,048}{4\,040} = 0.507$$

伯爵的铜板平衡吗？假设

$$H_0: p = 0.5$$

$$H_a: p \neq 0.5$$

要执行显著性检验，先把样本结果 $\hat{p} = 0.507$ 标示在 \hat{p} 的抽样分布上，这个抽样分布描述在原假设成立时 \hat{p} 值的变化情形。图 23.1 是图 22.3 的复制。图上显示出观测值 $\hat{p} = 0.507$ 离 0.5 不算远，并不能当做 p 的真正值不是 0.5 的合理证据。 P 值为 0.37，让我们的结论更明确。

假设布方伯爵掷铜板 1 000 次和 100 000 次，都得到同样的结果： $\hat{p} = 0.507$ 。当原假设为真时， \hat{p} 的抽样分布的平均数必定是 0.5，但它的标准差会随样本大小 n 的增加而减少。图 23.2 中画出了 $n = 1\,000$ ， $n = 4\,040$ 和 $n = 100\,000$ 时的三种抽样分布。图里面居中的那条曲线就是图 23.1 里的正态曲线，只是刻度改变，以便能够画得出 $n = 100\,000$ 时那条又高又窄的曲线。看看样本结果 $\hat{p} = 0.507$ 在三条曲线下的定位，你可以看出同一个结果几乎出乎预料地随样本大小而有不同。

P 值在 $n = 1\,000$ 时是 $P = 0.66$ ， $n = 4\,040$ 时是 $P = 0.37$ ，而在 $n = 100\,000$ 时就是 $P = 0.000\,009$ 了。想像一个重复投掷一个平衡铜板 1 000 次的情况。差不多有三分之二的时候，你得到的正面比例距为 0.5 的距离，会至少像布方的 0.507 距 0.5 的距离这么远。可是你如果是掷一个平衡铜板 100 000 次的话，几乎永远也不会（100 万回合中只会发生 9 次）得到这么稀奇的结果。

$\hat{p} = 0.507$ 这个结果如果是发生在掷 1 000 次铜板或 4 040 次铜板的状况下，并不是铜板不平衡的证据。但是如果发生在掷 100 000 次的时候，就成为铁一般的证据。

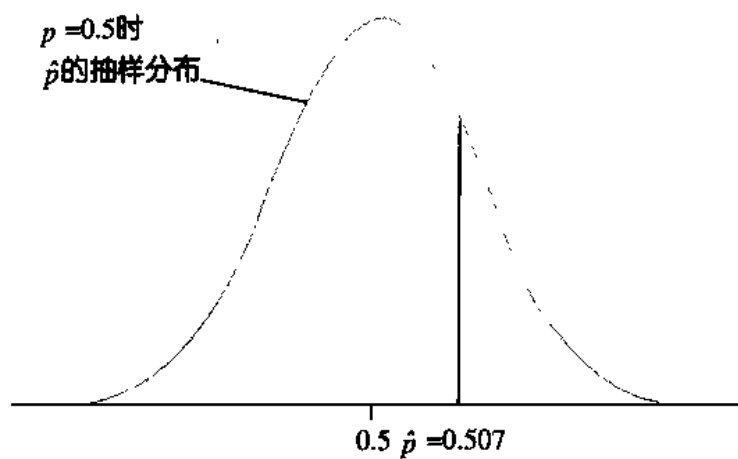


图 23.1 当铜板为平衡时，掷 4 040 次铜板所得正面比例的抽样分布。样本比例 0.507 不是不寻常的结果

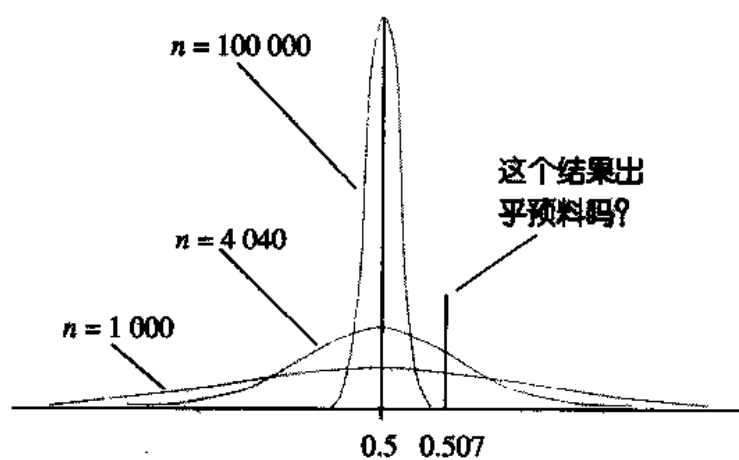


图 23.2 掷一个平衡铜板 1 000 次、4 040 次及 100 000 次所分别得到的正面比例的抽样分布。样本比例 0.507 在掷 1 000 次或 4 040 次的情况下很正常，但在掷 100 000 次的情况下就非常稀奇



留意光溜溜的 P 值

显著性检验的 P 值不仅和样本大小密切相关，也和总体真正值有关。若只报告 P 值，却不报告样本大小，也不提做为样本结果的统计量是什么，是很糟糕的做法。

置信区间的优点

例 1 告诉我们，要了解一项统计研究，不能只去看是否有统计显著性。光是知道样本比例是 $\hat{p}=0.507$ 就很有用处。你可以自己决定，这个值距 0.5 的差距，是否大到令你感兴趣。当然 $\hat{p}=0.507$ 并不是铜板真正的正面概率，而只是伯爵掷出来的机遇结果。所以置信区间会更有用，因为区间的宽度可以帮助我们真正的正面概率定位得更精确。

以下是真正的正面概率 p 的 95% 置信区间，分别对应例 1 中的三种样本大小。你可以验证一下，用第 21 章中教的方法可以得到这些结果。

投掷总次数	95% 置信区间
$n=1\,000$	0.507 ± 0.031 ，或 $0.476-0.538$
$n=4\,040$	0.507 ± 0.015 ，或 $0.492-0.522$
$n=100\,000$	0.507 ± 0.003 ，或 $0.504-0.510$

置信区间把我们对真正 p 值的了解(以 95% 的置信水平)明白表示出来。掷 1 000 次和掷 4 040 次所得到的区间都包含了 0.5 这个数字，所以我们不会去怀疑，铜板是否不平衡。可是掷 100 000 次的时候，我们却有信心真正的 P 值落在 0.504—0.510 之间。因此我们有信心 p 值不是 0.5。



统计学上的争议

应不应该禁止统计检验？

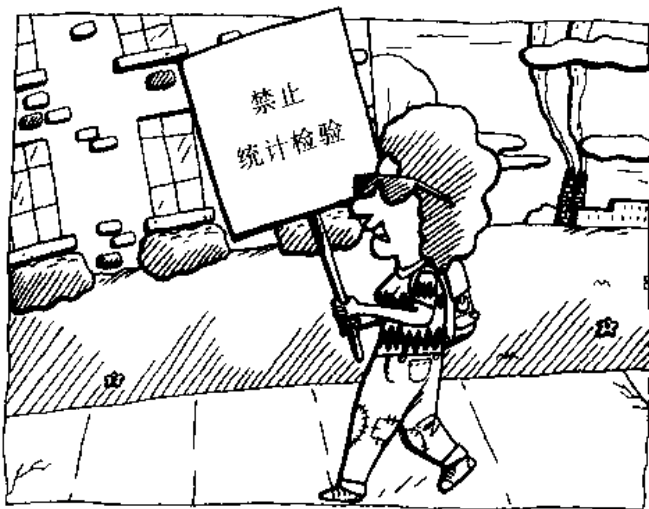
许多领域的研究都依赖显著性检验。习惯常常会导致过度依赖。哈佛大学一位因统计出名的杰出心理学家罗森塔尔(Robert Rosenthal)说：“我们当中有许多人所受的训练，是叫我们不要太仔细去看数据。你建立一项假设，决定用何种统计检验，然后执行该检验，如果你的结果有5%的统计显著性水平，你的假设就得到支持。否则就往抽屉一塞，再也不看这些数据。”

你听了应该会很吃惊。和“一定要问数据怎么来的”一样，“一定要把数据画图”也是我们的座右铭之一。心理学家通常很仔细地思考，数据要怎样产生。那怎么会有这么多心理学家，几乎看也不看数据呢？有些心理学家说，是因为显著性检验的惟我独尊，加上5%的显著水平被当做结果足够重要的魔术指标。尤其“结果要想发表，就必须有5%的显著水平”已成了惯例，所以研究者就养成了罗森塔尔形容的那种坏习惯。有5%的统计显著水平，成功；没有5%的统计显著性水平，失败。批评者说，对检验的限制这么大，解释错误的风险这么高且坏习惯这样根深蒂固，心理学专业期刊应该全面封杀显著性检验。

美国心理学会的回应，是针对统计推论指定了一个专门调查委员会。罗森塔尔为该委员会的共同主席之一。调查委员会不认为应该禁止统计检验。该会

做出的报告，事实上是如何执行优质统计工作的大纲。清楚定义你的总体；说明数据如何产生且尽可能用随机方法；描述你的变量和度量方法；说明样本大小以及是如何决定该大小的。如果有受试者半途退出或其他实际问题，报告中也要提。一旦你已搜集了数据，在计算任何统计量之前，先看看数据。看看计算结果是否合理。要认清“根据非随机化设计来做有关因果的结论，是非常冒险的做法”。

其他还有很多，不过以下是有关统计检验的重要关键：“很难想像在怎样的情况下，光是做接受或拒绝的决定，会优于报告出 p 值，或者优于置信区间……报告 p 值时也一定要同时提出‘效应大小’的估计值。”要禁止做检验？“虽然禁止可以防止滥用，但是委员会认为反例更多，足够支持检验继续存在。”





• 提出置信区间

置信区间提供的信息比检验多，因为置信区间实际上估计了总体参数的值。而且置信区间也比较容易解释。因此，好的做法是尽可能提出置信区间。

“5%的显著水平”并非魔术指标

显著性检验的目的，是要描述样本所提供不利原假设的证据有多强。 P 值就是在做这件事。但是要证明原假设不正确， P 值要多小，才能令人信服呢？这点主要根据两种状况来决定：

- H_0 的可信度有多高？如果 H_0 所代表的假设，是人们多年来一直相信的，就需要很强的证据（小的 P 值）才能说服他们。
- 拒绝 H_0 的结果是什么？如果拒绝 H_0 而肯定 H_A ，代表要花很多钱把产品包装改换成另一种，你就需要有很强的证据，显示新包装一定会增加销售量。

这两种标准都有一点主观。不同的人常会想用不同的显著水平。报告 P 值，可以让我们各自决定证据是不是够充分。用统计的人常会强调某几个标准的显著水平，比如 10%、5% 和 1%。举例来说，法庭在一些歧视案件里常用 5% 当作标准。会把重点放在这几个值，反映出做统计的人仍在使用临界值，而尚未进入电脑软件的时代。5% 的显著水平 ($\alpha = 0.05$) 尤其常用。在“显著”和“不显著”之间并没有清楚的界线，只是在 P 值愈来愈小时，我们就有愈来愈强的证据而已。0.049 和 0.051 这两个 P 值，并没有多少实质的差别。

把 $P \leq 0.05$ 当作“显著水平”的全球性标准，一点道理也没有。

提防刻意寻找的显著性

统计显著性的意义应该是：你找到了你在寻求的效应。假如，你先决定你在寻求什么效应，设计研究来找这个效应，再用显著性检验



来估量你得到的证据，那么统计显著性背后的论据就可以充分发挥。假如不是这样的结构，显著性的意义可能不大。我们在本章开头时看到一个例子：先检视 6 700 种共同基金，找出刚好获利最高的那一只，然后来问是否这只基金“显著”优于平均获利，这样做一点道理也没有。以下是另一个例子。

例 2 预测受培训者的日后表现

接受管理方面培训的人，有的最后变成管理者，有的花了很多钱训练之后，仍然没有长进，只得离开公司。你想要知道是什么因素造成这样大的区别吗？你有一大堆过去受培训者的资料，包括他们的个性和目标、大学时学了什么以及表现如何，甚至他们的家庭背景和嗜好。利用统计软件，可以轻易对这一大堆变量执行很多个显著性检验，来看看哪些变量最能够预测未来的成功。啊哈！结果你发现，和被淘汰的人比起来，未来的管理者明显有较多的人，有城市或郊区的成长背景，以及专业领域的学士学位。

在你决定以后要根据这些发现来求才之前，先要记住，有 5% 的显著性水平的结果，即使在 H_0 为真的时候，长期下来 100 次当中也会发生 5 次。当你做很多个显著水平为 5% 的检验时，你会预期其中有几个单单因为机遇就有显著性。做一个检定并达到 $\alpha = 0.05$ 的，是你已有所发现的好证据。做了好几十个检验，然后有一两次达到标准，可就不是什么证据了。

在共同基金例子里，我们先挑出最好的，然后又去对它做检验，好似我们并没有先把它挑出来一样。在例 2 当中，我们检验了每种可能，再挑出最具显著性的。这两种错误示范，会混淆“数据的探索分析”及“正式的统计推论”这两个角色。

在资料里搜寻可能的形态当然是合理的。探索性数据分析 (exploratory data analysis) 是统计中重要的一部分。但是，如果你已经成功的在资料里找到突出的效应，正式推论的论据就不再适用。补救方法很清楚。一旦你有假设后，设计一个研究来找寻你认为有的这个特定效应。如果这个研究结果有统计显著性，就有真正的证据了。



网络寻奇

美国心理学会的“统计推论专门调查委员会”所做的报告，对如何聪明的应用统计推论，提供了很好的简介。这份报告是在 1999 年的《美国心理学家》(*American Psychologist*) 期刊上。你可以上网，在 www.apa.org/journals/amp.html 网站上这份期刊的“selected articles”清单中找到这篇名为 *Statistical Methods in Psychology Journals: Guidelines and Explanations* 的报告。



本章重点摘要

统计推论的应用范围不如探索性数据分析广泛。任何推论方法都只能在正确的设置之下应用，尤其要符合随机样本或随机实验的设计。

了解置信区间和统计显著性的意义，有助于避免不恰当的结论。增加观测值的数目对置信区间有很直接的影响，因为在同样的置信水平之下区间会变短。即使总体的真正情况维持不变，观测值变多，通常会把检验的 P 值变小，使得检验比置信区间要难解释些。样本很大时，即使结果的 P 值很小，也不见得有实际上的意义；而样本小时，总体的重要真实情况却可能达不到统计显著性的标准。要避免用固定的显著水平，例如 5% 的显著水平之类来做决定。



第 23 章 习题

23.1 电视台民意调查。某电视台的新闻节目做了一项打电话回应民意调查，内容是关于市政当局拟议中的禁止拥有枪支方案。在打进去的 2372 个电话中，有 1921 个反对禁止持枪。该电视台遵循一般做法，发表了置信叙述：“第 13 频道的意向调查样本中，有 81% 反对禁止拥枪。我们有 95% 信心，市民中反对禁止持枪的真正比例，会在样本结果的正负 1.6% 范围内。”置信区间的计算是正确的，然而结论的理由不充分。为什么？

23.2 美国总统的年龄。乔在写一篇有关美国总统背景的报告。他查了所有 43 位总统入主白宫时的年龄。因为乔曾修了一门统计课，他就用这 43 个数字，造了一个历任总统平均年龄的 95% 置信区间。这样做一点也行不通。为什么？

23.3 谁会赢？一项在临近选举时进行的民意调查显示，选民当中支持甲而不支持乙的占了 52%。在 95% 置信水平之下，该民意调查结果的误差界限为正负 3 个百分点。针对该项调查结果发布的新闻报导中说，选举结果太接近，尚无法预测谁会得胜。为什么？

23.4 有钱父母的影响到哪里为止？小孩会受多少教育，和父母的财力及社会地位有密切关系。用社会学术语来说，这叫“社会经济地位” (SES, socioeconomic status)。但是已经大学毕业的小孩会不会再深造，就不太受父母的 SES 影响。有一项研究检视了大学毕业生参加商学院、法学院及其他领域的研究所入学测验的状况。父母的 SES 对参加法学院的 LSAT 测验的影响为“既无统计显著性，又很微小”。

(a) “无统计显著性”是什么意思？

(b) 为什么除了无统计显著性以外，“影响很小”也是重要资讯？

23.5 寻找特异功能。一位研究者寻求特异功能 (ESP) 存在之证据，检测了 500 个人，其中有 4 个人的测验结果显著好于瞎猜 ($P < 0.01$)。

(a) 做出此 4 人有 ESP 的结论是否恰当？说明你的理由。



(b) 如果研究者要检验这 4 个人当中是否有人有 ESP, 现在该怎么做?

23.6 比较包装设计。一家公司为了比较某名牌洗涤剂的两种包装设计, 把两种设计的瓶子都放在各市场的货架上。收银台对超过 5 000 瓶售出的洗衣精的扫描资料显示, 买设计 A 的人比设计 B 的人多。其差异具统计显著性 ($P=0.02$)。我们能不能下结论说, 消费者对设计 A 要满意得多? 要说明理由。

23.7 非洲的色盲探讨。一位人类学家察觉, 在靠打猎及采集为生的社会中, 似乎色盲状况没有农业社会中严重。他在非洲两群人中各检测了一些成人, 两群人分属以上两种社会。在打猎及采集那群人口中, 色盲者比例明显低于 ($P<0.05$) 另一群人口。你还需要哪些信息, 才能决定是否接受该有关色盲者比例的声明?

23.8 东南亚的血型探讨。要判断两群人是否分属不同群体的方法之一, 是比较该两群人的血型分布。一位人类学家在马来西亚中部的不同部落之间, 找到了人类主要血型 (A、B、AB、O) 比例的显著差异 ($P=0.01$)。你还需要哪些信息, 才能同意这些部落的确分属不同群体?

23.9 为什么要寻求显著性。在被问到为什么统计显著性在研究报告中出现如此频繁时, 一位学生回答: “因为声明结果有统计显著性就是在说, 该结果不能轻易只用机遇变异来解释。”你觉得这个说法是否基本上正确? 要说明理由。

23.10 显著性对什么样的问题有用? 以下的问题当中, 哪些可以用显著性检验来回答?

- (a) 样本或实验的设计是否合适?
- (b) 观察到的效应是否因为机遇而产生?
- (c) 观察到的效应是否重要?

23.11 精神分裂者如何区分? 心理学家曾对一个精神分裂者样本和一个非精神分裂者样本度量了 77 个变量的值, 他们分别用了 77 个显著性检验, 来比较这两个样本。其中有 2 个检验有 5% 的统计显著性水平。假设该 77 个变量在成人总体的精神分裂者和非精神分裂者之



间，事实上都没有差异。也就是说，所有 77 项原假设均为真。

- (a) 其中某个特定的检验会显示出具有 5% 的显著水平的差异的概率是多少？
- (b) 为什么 77 个检验中有 2 个有 5% 的显著水平这件事，一点也不出人意料？

23.12 为什么大一点的样本比较好？统计学家比较喜欢大一些的样本。简短描述一下，增加样本大小(或增加实验中的受试者人数)会对以下项目发生什么影响。

- (a) 95% 置信区间的误差界限。
- (b) 当 H_0 不正确，而所有其他有关总体的事实均无改变时检验的 P 值。

23.13 这样可信吗？你计划针对一种现在还没有疫苗可对付的病毒，检验一种新疫苗的有效性。因为这种病并不严重，所以你会让 100 位自愿受试者暴露在有这种病毒的环境之下。在一段时间之后，你会记录每一位受试者是否受到感染。

- (a) 请说明你会怎么样设计实验，来用这 100 位志愿者检验疫苗。要包括设计实验时的所有重要细节(但不必实际执行随机化的部分)。
- (b) 你希望能显示出疫苗比安慰剂有效。写出 H_0 和 H_a 。(请注意这个检验是在比较 2 个总体比例。)
- (c) 实验结果的 P 值是 0.25。详细说明这是什么意思。
- (d) 你的研究伙伴认为证据不够，不足以支持正式使用新疫苗的建议。你同意吗？

以下习题会用到 21 章和 22 章中可略过部分所提到的方法。

23.14 我们的运动员有没有毕业？让我们回到习题 22.12 中的研究，研究结果发现，147 位进入某大型大学就读的运动员中，有 95 人在 6 年内毕业。运动员的毕业比例显著低于 ($P=0.053$) 所有学生的 70.7% 毕业比例。如果造一个运动员毕业比例的 95% 置信区间，可能会提供更多信息。请算出这个区间。

23.15 我们想要有钱。回到习题 22.13 中的研究。该研究发现，在



美国某州立大学的 200 位入学新生中,有 132 人认为“很有钱”很重要。这项结果和全美国一年级新生中 73% 有这种想法的比例,有显著差别 ($P=0.026$)。如果造一个该州立大学入学新生认为有钱重要的比例的 95% 置信区间,可能会提供更多信息。请算出这个区间。

23.16 有统计显著性吗? 根据若干年期间数千名学生的资料,某大城市的高中学生当中,有 895 通过了一项能力测验,通过这项测验是要拿毕业文凭的必备条件之一。有些改革派认为一种新的数学课程可以增加通过比例。一个 1 000 位学生的随机样本修习了新课程。学校董事会认为,进步情况要达到 5% 的统计显著性水平,才准备对全体学生全面实施新课程。假设 p 代表所有学新课程的学生中会通过测验的比例,我们就应该检验

$$H_0: p = 0.89$$

$$H_a: p > 0.89$$

- (a) 假设样本的 1 000 名学生中,有 906 人通过测验,请证实这项结果达不到 5% 的统计显著性水平。(用 22 章例 3 的方法。)
- (b) 假设样本的 1 000 名学生中,有 907 人通过测验,请证实这结果有 5% 的统计显著性水平。
- (c) 1 000 次当中有 906 次成功,还是有 907 次成功,实际上有什么差别吗? 你对于“有 5% 的统计显著性水平”的重要性有什么看法?

23.17 我们喜欢置信区间。上一题习题当中比较了对于所有学生中会通过能力测验者的比例 p 的显著性检验,两项检验所根据的数据,是 1 000 名学生 SRS 中,分别有 906 人或 907 人通过测验。对两项结果各造一个 95% 置信区间。从这两个区间可以看出,我们对真正的 p 值多公没把握,以及两项样本结果之间的差异多么小。

.

第 24 章

双向表及卡方检验*

女性大学教授

笔者任教的普度大学属于“十大”(Big Ten)系统,该校的重点在于工程、科学及技术领域。在 1998—1999 年的学年度,普度共有 1 621 位教授,其中 335 位为女性。女性所占比例只稍高于 20%,也就是大约每 5 位教授中有 1 位是女性。光看这些数字,无法看出女性在教授中的地位。就如一向的处理方式,我们必须检视数个变量之间的关系,而不能只看性别。举例来说,在人文学科中,女性教授的比例就比农业学系的要高。

我们来看看性别和职称之间的关系,后者是对于教授极为重要的

* 此章为选读。



一个变量。教授通常从助理教授 (assistant professor) 做起, 之后升到副教授 (associate professor) 并取得终身职, 最后才达到正教授 (full professor) 的位置。大学运作的主要决策者差不多都是正教授。以下是把普度大学 1 621 位教授依性别以及职级分类的双向表 (two-way table):

	女性	男性	总数
助理教授	126	213	339
副教授	149	411	560
正教授	60	662	722
总数	335	1 286	1 621

女性在教授中的地位, 从这个表来看就清楚多了。我们顺着职位从低往高爬时, 男性的人数愈来愈多, 女性的人数却愈来愈少。比例透露的信息会比计数更清楚, 所以我们来计算一些百分比。助理教授中超过 37% 是女性, 然而女性在副教授中只占约 27%, 正教授中只占约 8%。女性教授在最高职称中所占的比例超低。

表列数字只能呈现事实但不能解释事实。一般来说助理教授要花 6 年的时间才能升到副教授, 然后再花好些年才能升到正教授。有可能许多女性教授是在近十年内才加入教授阵容, 职称低是因为还年轻。或者也许女性比较不容易升等, 这就比较严重。必须要多检视些数据, 才能判断哪种解释比较合适。



双向表

大学教授的职级和性别都属于类别变量。也就是说, 这些变量会被分类, 但是并没有数值可以让我们用来画散布图、计算相关系数或回归直线以描述相关关系。要显示两个类别变量之间的相关关系, 可以用像普度教授的职级及性别表那样的双向表。职级是列变量 (row variable), 因为表中每一列代表教授的一种职级。性别是行变量



(column variable), 因为每一行代表一种性别。表中的数字是属于每一种职级对性别的组合所含教授人数。虽然职级和性别都属于类别变量, 职级却有由低到高的自然顺序。表中列的顺序反映了该类别的顺序。

怎么样可以最有效的从这当中掌握信息呢? 首先, 分别检视每个变量的分布。类别变量的分布告诉我们每个结果发生的频繁程度。表最右的“总数”栏, 列出的数字是每一列的总数。这些列总和(row totals)提供了所有男、女教授的职称分布。表底端那一列“总数”, 提供的是所有教授的性别分布。通常用百分比表示这些分布会更清楚。我们可以把性别分布表示成:

$$\text{女性百分比} = \frac{335}{1\ 621} = 0.207 = 20.7\%$$

$$\text{男性百分比} = \frac{1\ 286}{1\ 621} = 0.793 = 79.3\%$$

双向表所包含的信息, 不只是单独的职称分布和性别分布。因为职级和性别之间有何关系, 没有办法从个别的分布当中找出来, 一定要用到整个双向表。要描述类别变量之间的相关关系, 可根据表中所给的计数, 计算出适当的百分比。

例 1 教授的职称和性别

因为性别只有两种, 我们可以通过比较女性在三种职称所占之百分比, 来检视性别和职称之间的关系。

助理教授	副教授	正教授
$\frac{126}{339} = 37.2\%$	$\frac{149}{560} = 26.6\%$	$\frac{60}{722} = 8.3\%$

从百分比可以清楚看见, 女性在较高的职称中较少见。这就是性别和职称间相关的本质。



用双向表的时候必须计算许多百分比。以下这一招可以帮你决定，用哪些分数才可以算出你要的百分比。要问自己：“我要的百分比是哪一个整体的百分比？”该整体的计数，就是你要算百分比时所用分数的分母。在例1当中我们要找的是，每一个职称中女性所占的百分比，所以每一个职称的计数就成为分母。

辛浦森悖论

就像数量变量的情形一样，潜在变量的效应有可能改变、甚至倒转两个类别变量间相关的方向。让我们继续以性别和高等教育的主题当做例子。数字是为了简化问题而造出来的，但是仍说明了常常在真实数据中出现的现象。

例2 入学审核有性别歧视？

某大学只有两个科系：一是电机工程、另一是英文。申请入学并不容易，而妇女委员会怀疑，审核过程有歧视女性的嫌疑。委员会从学校得到以下资料，是列出所有申请者的性别和审核结果的双向表：

	男性	女性
通过	35	20
不通过	45	40
总和	80	60

资料的确显示出：申请者的性别和是否获得审核通过之间，似有某种相关关系。为了要更精确地描述这个相关性，我们来计算一些百分比。

$$\text{男性申请者通过的百分比} = \frac{35}{80} = 44\%$$

$$\text{女性申请者通过的百分比} = \frac{20}{60} = 33\%$$

啊哈！男性几乎有一半申请成功，女性通过的却只有三分之一。



该大学的答复是，虽然资料显示的相关性是正确的，却不是因为性别歧视造成的。为了替自己辩护，学校制作了三向表(three way table)，将每个申请者依性别、申请结果及所申请的主修三个变量分类。我们将三向表以数个并列的双向表形式呈现，每个双向表对应第三个变量的一个值。在这个例子中共有两个双向表，各对应一个主修。

电机工程			英文		
	男性	女性		男性	女性
通过	30	10	通过	5	10
不通过	30	10	不通过	15	30
总和	60	20	总和	20	40

“辛浦森悖论”创造者的工作情境



“不，我和辛浦森家庭里霸子他爹无关。”

你可以先检查一下，把这两个表里面同样位置的值加起来，就和先前的双向表里的一样。该大学只是依申请系别把双向表的资料拆成了两个表。现在我们可以看到，电机系正好收了申请学生的半数，对男生、女生都如此，而英文系对男、女申请者各收四分之一。不管哪个系，性别和通过与否都没有相关性。

这是怎么回事呢？分别来看，任何系的资料之间都没有相关性，放在一起，却有很强的相关性。我们来仔细看看资料。英文系比较难申请，但申请该系的大部分是女生。电机系较容易申请，而申请的主要是男生。英文系的申请者有 40 个女生，20 个男生；而电机系是 60 个男生，20 个女生。最初的双向表没有考虑到各系之间的差别，因此会误导。这是辛浦森悖论(Simpson's paradox)的一个例子。

• 辛浦森悖论

辛浦森悖论指出，在几组值中都显示出的相关关系或比较，有可能在数据合并成一组时全都消失甚至倒转方向。



当潜在变量存在时，观察到的相关关系有可能是误导的，而辛浦森悖论只是这项事实的一种极端形式。要记住第15章的警告：注意潜在变量。

例3 对抵押贷款对象有歧视？

研究银行的房屋抵押贷款申请，就可以看出结果和种族大有关系：黑人申请者被拒的比例高于白人申请者。在首都华盛顿地区有一件诉讼案声称：某家银行拒绝了17.5%的黑人贷款申请，却只拒绝了3.3%的白人。

银行答复：潜在变量可以解释拒绝比例的差距。比起白人(平均来说)，黑人收入较低、信用记录较差、较少有稳定工作。而与种族因素不同的是，这些事实是拒绝抵押贷款的合法理由。银行说，因为这些潜在变量和种族因素纠缠不清，所以他们拒绝的黑人申请者比例才会较高。想想辛浦森悖论就知道，如果以收入和信用记录相同的人来看的话，银行通过黑人申请者的比例可能还高过白人呢！

到底谁对？双方都会聘请统计学家来评估潜在变量的效用大小，而法院终会做出决定。

双向表的推论

我们常常搜集数据，并列出双向表，来探讨两个类别变量之间是否有关系。要检视样本数据很容易：算出百分比，再来看看列变量和行变量之间有没有相关关系。样本显示出的相关关系，是不是就证明了整个总体中这两个变量有相关关系呢？还是样本中的相关关系很容易只因为随机抽样的巧合就发生了呢？这是显著性检验的问题。



例 4 治疗可卡因瘾

对可卡因上瘾的人需要靠它来得到快感。也许给这些人抗忧郁的药，能帮助他们戒掉可卡因瘾。有一项历时三年的研究，把一种叫做去郁敏的抗抑郁剂和锂盐(治疗可卡因瘾患者的标准用药)以及安慰剂做了比较。受试对象是 72 位长期使用可卡因但是想要戒掉的人。每一种处理都各随机指派了 24 个人。以下是在研究期间，成功做到不碰可卡因的人数及百分比：

组	处理	人数	成功人数	百分比(%)
1	去郁敏	24	14	58.3
2	锂盐	24	6	25.0
3	安慰剂	24	4	16.7

受试者中成功做到不碰可卡因的样本比例有很大差别。去郁敏尤其比锂盐或安慰剂的效果好得多。图 24.1 的柱状图呈现了这个差别。这是不是合理证据，显示了在所有可卡因瘾君子的总体当中，处理和结果的确有相关关系？

可以对这个问题提供答案的检验，会以一个双向表为基础。以下就是对应例 4 中数据的双向表：

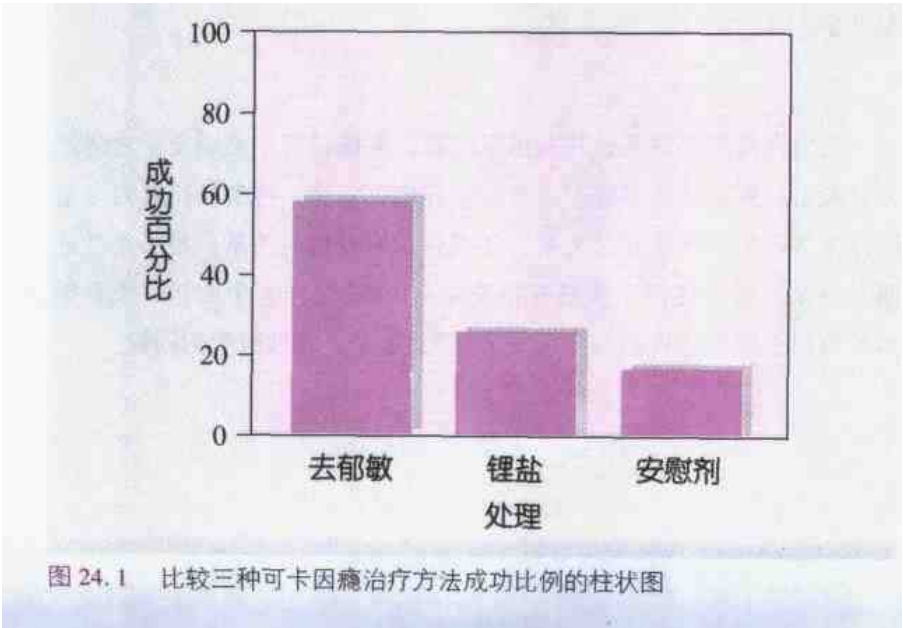


图 24.1 比较三种可卡因瘾治疗方法成功比例的柱状图



	成功	失败	总数
去郁敏	14	10	24
锂盐	6	18	24
安慰剂	4	20	24
总数	24	48	72

我们的原假设照例是说各个处理都没有效应。也就是说，瘾君子在三种处理之下的表现都一样。样本中显示出来的差异不过是机遇造成的。我们的原假设是：

H_0 ：在所有可卡因成瘾患者的总体当中，处理和戒瘾成功之间并没有相关关系。

把这个原假设用总体参数表示会有一点复杂，所以我们用这样的叙述就可以了。备择假设说的只是“是的，瘾君子所接受的处理，和是否能成功戒除可卡因，二者之间的确有相关关系”。备择假设并没有指明相关关系的本质。比如说，备择假设并没有说“使用去郁敏的患者，比起使用锂盐或安慰剂的患者，更有机会戒瘾成功”。

要检验 H_0 ，我们会把双向表中已观察到的计数和预期计数* (expected count)，做比较。预期计数是当 H_0 为真时，我们会预期的计数(除了随机变异外)。如果观察到的计数和预期计数相差很大，就是不利于 H_0 的证据。我们可以猜得出可卡因研究中的预期计数。以全部受试者来说，72 人中有 24 人成功。这代表整体成功率是三分之一，因为 $24/72$ 等于三分之一。如果原假设为真，各处理之间就没有差别。所以我们预期见到每一组当中有三分之一的人成功。每组里面有 24 个人，所以我们预期见到每组当中有 8 人成功，16 人失败。如果每个处理组的人数不尽相同，则预期计数也会不同，即使我们仍然预期每一组有相同的成功比例。幸好有个公式可以帮我们轻易算出预期计数，详列于下。

• 预期计数

H_0 为真时，双向表中任一格的**预期计数**(expected count)为：

$$\text{预期计数} = \frac{\text{列总和} \times \text{行总和}}{\text{表总和}}$$

其他检验

有的检验对付的是比“没有相关关系”更明确的假设。将某些人依社会地位分等级，等上十年，再把同样一批人重新分等级。列变量和行变量分别是两个不同时间的等级。我们可以检验社会地位的整体分布并没有改变的假设。也可以检验是否社会地位上升的人和下降的人比例差不多。某些统计检验可以处理这种以及其他种类的假设。

* 译注：通常 expected 都翻译成“期望”，但此处“预期”比“期望”更为贴切。



试用这个公式看看。例如去郁敏组的预期成功计数为：

$$\begin{aligned}\text{预期计数} &= \frac{\text{第一列总和} \times \text{第一行总和}}{\text{表总和}} \\ &= \frac{(24)(24)}{72} = 8\end{aligned}$$

如果原假设说的处理之间无差异为真，我们会预期去郁敏组的 24 人中有 8 人成功。这和我们之前猜的一样。

卡方检验

想知道数据是否提供了不利于“没有关系”的原假设的证据，我们得把双向表里的计数，和假设如真的没有关系时我们会预期的计数做比较。假如观察到的计数和预期计数相差很多，我们就得到了想找的证据。这个检验用了一项统计量，来度量观察到的与预期的计数到底相差多少？

• 卡方统计量

卡方统计量 (Chi-square statistic) 度量出双向表中观察到的计数和预期计数之间的差距。统计量的公式是：

$$\chi^2 = \sum \frac{(\text{观察到的计数} - \text{预期的计数})^2}{\text{预期计数}}$$

\sum 这个符号代表“对应表里面每一格加总起来”。

卡方统计量是许多项数字的和，每一项对应表里面的一格。在可卡因的例子中，去郁敏组当中有 14 人成功，而这格的预期计数是 8。所以卡方统计量中对应该格的这项数字是：

$$\begin{aligned}\frac{(\text{观察到的计数} - \text{预期的计数})^2}{\text{预期计数}} &= \frac{(14 - 8)^2}{8} \\ &= \frac{36}{8} = 4.5\end{aligned}$$

**例 5 可卡因研究**

以下是可卡因研究中的观察计数和预期计数，两者并列在同一个表当中：

	观察到的		预期	
	成功人数	失败人数	成功人数	失败人数
去郁敏	14	10	8	16
锂盐	6	18	8	16
安慰剂	4	20	8	16

现在我们可以算出卡方统计量，只要把双向表当中 6 个格子所对应的项加起来即可：

$$\begin{aligned}\chi^2 &= \frac{(14-8)^2}{8} + \frac{(10-16)^2}{16} + \frac{(6-8)^2}{8} + \frac{(18-16)^2}{16} + \frac{(4-8)^2}{8} + \frac{(20-16)^2}{16} \\ &= 4.50 + 2.25 + 0.50 + 0.25 + 2.00 + 1.00 = 10.50\end{aligned}$$

因为 χ^2 度量的是观察到的计数距 H_0 为真时的预期计数的差距，所以它的值如果偏大，就是不利于 H_0 的证据。 $\chi^2 = 10.5$ 算不算大呢？你知道该怎么处理：把 10.5 这个观察值和抽样分布做个比较， χ^2 的抽样分布会显示出在原假设为真时， χ^2 的值会有怎样的变化。这项抽样分布不是正态分布，而是右偏分布：又因为 χ^2 的值不可能为负，所以只含大于 0 的值。还有，对应不同大小的双向表，抽样分布也会不同。实际状况如下所述。

- 卡方分布**

当“无相关关系”的原假设为真时，卡方统计量 χ^2 的抽样分布就叫做卡方分布 (chisquare distribution)。

卡方分布指一整族分布，而这个分布只有正值且为右偏。特定的卡方分布是由它的自由度 (df, degrees of freedom) 决定的。



有 r 列和 c 行的双向表所对应的卡方检定，用的是自由度为 $(r-1)(c-1)$ 的卡方分布之临界值。

图 24.2 中画出了三种卡方分布的密度曲线。当自由度增加时，密度曲线的偏斜程度会减小，而较大值出现的可能性加大。我们没有办法用纸笔计算出卡方曲线底下的面积来找出 P 值，但可以用软件来算。而表 24.1 是条捷径，它列出了在不同的显著水平下，卡方统计量 χ^2 的值至少要多大，才能使结果有统计显著性。这虽然不如实际找出 P 值那么好，但是通常也够好了。每一种自由度在表中对应不同的列。比如说我们可以从表中查到，自由度为 3 的卡方统计量，如果值大于 7.81，则有 5% 的统计显著性水平；如果值大于 11.34，则有 1% 的统计显著性水平。

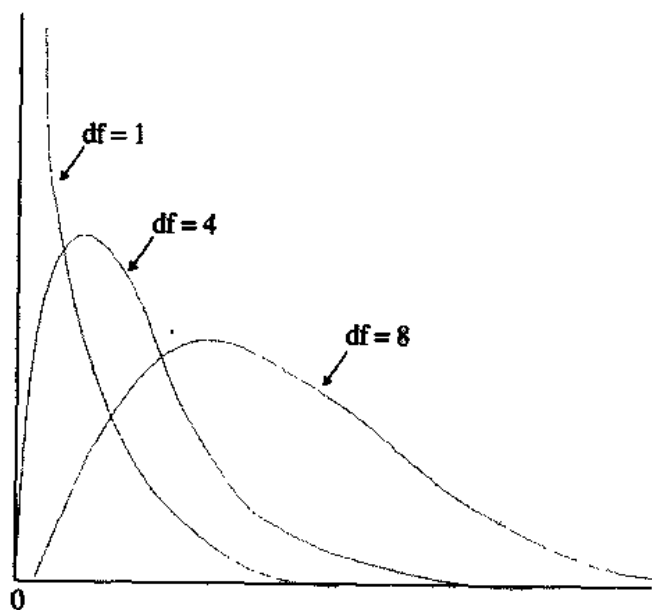


图 24.2 卡方分布族中三个成员的密度曲线。卡方统计的抽样分布属于此族

例 6 可卡因研究总结

我们已经见到，去郁敏和锂盐及安慰剂比起来，明显有较多的成功案例，而失败案例较少。经过比较观察计数和预期计数之后，得到卡方统计量 $\chi^2 = 10.5$ 。最后只剩下评估统计显著性这一步。

可卡因研究的双向表由三种处理和两种结果所构成，共有 3 列、2 行。也就是说 $r=3$ 及 $c=2$ 。因此卡方统计量的自由度为：

$$(r-1)(c-1) = (3-1)(2-1) = (2)(1) = 2$$

表 24.1 要在显著水平为 α 时有统计显著性，卡方统计量的值必须大于 α 所对应的那一行的值

df	显著水平 α						
	0.25	0.21	0.15	0.10	0.05	0.01	0.001
1	1.32	1.64	2.07	2.71	3.84	6.63	10.83
2	2.77	3.22	3.79	4.61	5.99	9.21	13.82
3	4.11	4.64	5.32	6.25	7.81	11.34	16.27
4	5.39	5.99	6.74	7.78	9.49	13.28	18.47
5	6.63	7.29	8.12	9.24	11.07	15.09	20.51
6	7.84	8.56	9.45	10.64	12.59	16.81	22.46
7	9.04	9.80	10.75	12.02	14.07	18.48	24.32
8	10.22	11.03	12.03	13.36	15.51	20.09	26.12
9	11.39	12.24	13.29	14.68	16.92	21.67	27.88

在表 24.1 中查自由度 $df=2$ 的那一行。我们会看到， $\chi^2 = 10.5$ 大于 $\alpha = 0.01$ 的显著水平所对应的临界值 9.21，但是小于 $\alpha = 0.001$ 的显著水平的临界值 13.82。可卡因研究显示，处理和成功之间有具统计显著性的相关关系 ($P < 0.01$)。

显著性检验只是说，我们看到有力证据，证明处理和成功之间有某种相关关系。要检视双向表才能找到相关关系的本质：去郁敏的效果优于其他处理。



如何应用卡方检验

就像我们对总体比例做的检验一样，卡方检验也用了一些近似结果，我们的观测值愈多，结果就愈精确。以下是何时适合用此检验的大致规则。

• 应用卡方检验所需的每格计数下限

当预期计数小于 5 的格子所占比例不超过 20%，而且每一格的预期计数都至少是 1 时，就可安心使用卡方检验。

可卡因研究轻易就通过这项资格测验：它所有的预期格计数不是 8 就是 16。我们以下面这个讨论如何检视双向表的例子，作为本章的结尾。

例 7 容易生气的人是否较易得心脏病？

容易生气的人似乎较可能得心脏病。这是一项历时约四年的研究所得到的结论，该项研究追踪了三地共 12 986 人的随机样本。所有受试者在研究开始之初都没有心脏疾病，他们都接受了斯皮尔伯格发怒量表的检测，该量表用来度量一个人有多容易发怒。以下是样本中血压正常的 8 474 人的数据。CHD 代表冠状动脉心脏病 (coronary heart disease)，包括曾有过心脏病猝发的人，以及需要做治疗的心脏病患者都属此类。

	易怒指标		
	低	中	高
样本大小	3 110	4 731	633
CHD 计数	53	110	27
CHD 百分比 (%)	1.7	2.3	4.3

我们可以看到明显的趋势：易怒指标愈高，心脏病百分比就愈高。发怒和心脏病之间的这个相关关系，是否有统计显著性？

第一步要先把数据用双向表表示，并加进没有心脏病的人的计数。我们也加入了在算预期计数时需要用到的列总数及行总数。

	不易怒	中度易怒	高度易怒	总数
CHD	53	110	27	190
无 CHD	3 057	4 621	606	8 284
总数	3 110	4 731	633	8 474

现在我们可以遵循显著性检验的步骤了，这是在第 22 章中早已熟悉的。

假设。卡方检验是检验下列假设：

H_0 : 易怒和 CHD 之间没有相关关系

H_a : 易怒和 CHD 之间有相关关系

抽样分布。我们会看到，所有的预期格计数都比 5 大，所以我们可以放心使用卡方检验，是否易怒对应 CHD 的双向表有 2 列及 3 行。我们将从自由度为 $df = (2 - 1)(3 - 1) = 2$ 的卡方分布中找临界值。

数据。先算出预期格计数。举例来说，高度易怒且有 CHD 的那一格，其预期计数是：

$$\begin{aligned} \text{预期计数} &= \frac{\text{第一列总和} \times \text{第三行总和}}{\text{表总和}} \\ &= \frac{(190)(633)}{8\,474} = 14.19 \end{aligned}$$

以下是同时列出所有观察到的以及预期计数的表：

	观察到的			预期		
	低	中	高	低	中	高
CHD	53	110	27	69.73	106.08	14.19
无 CHD	3 057	4 621	606	3 040.27	4 624.92	618.81



检视这些计数就会看到，高度易怒组的 CHD 人数比预期的高，而不易怒那组的 CHD 人数比预期的低。这个结果和例 7 当中所列出的百分比是一致的。卡方统计量为：

$$\begin{aligned}\chi^2 &= \frac{(53 - 69.73)^2}{69.73} + \frac{(110 - 106.08)^2}{106.08} + \frac{(27 - 14.19)^2}{14.19} \\ &\quad + \frac{(3\,057 - 3\,040.27)^2}{3\,040.27} + \frac{(4\,621 - 4\,624.92)^2}{4\,624.92} + \frac{(606 - 618.81)^2}{618.81} \\ &= 4.014 + 0.145 + 11.557 + 0.092 + 0.003 + 0.265 = 16.077\end{aligned}$$

实际应用时可利用统计软件做所有的计算。看一下相加得出 χ^2 的那 6 个数字。 χ^2 的值主要是由其中一格“贡献”的：高过预期的高度易怒者的 CHD 人数。

有统计显著性吗？在表 24.1 中查 $df=2$ 那一列。从数据算出的卡方统计量 χ^2 为 16.077，比对应 $\alpha=0.001$ 的临界值 13.82 要大。我们获得了具高度统计显著性的证据 ($P < 0.001$) 显示易怒和心脏病的确有关。统计软件可以算出实际的 P 值，答案是 $P=0.0003$ 。

我们能不能下结论说，容易发怒会导致心脏病呢？这是一项观测研究，而不是实验。如果有些潜在变量和易怒相交叉，也不会叫人意外。比如说，易怒的人比起其他人来，比较有可能是既喝酒又抽烟的男性。研究报告中用了高等统计，对三组不同发怒程度的人之间的许多种差异做了调整。经过调整之后， P 值从 $P=0.0003$ 上升到了 $P=0.002$ ，因为潜在变量可以对心脏病做部分解释。这个结果仍然是有相关关系的合理证据。因为研究是从没有 CHD 的人的随机样本开始，然后追踪观察这些人，还有因为许多潜在变量都经过度量并加以解释，所以研究结果的确对因果提供了部分证据。下一步也许应该做个实验，看易怒的人可以如何做改变，且这样是不是能减低他们得心脏病的风险？



本章重点摘要

类别变量把个体归类到不同组。要想呈现两个类别变量之间的相关关系，就用一个包含各组计数的双向表。我们通过比较某些特定的百分比，来描述类别变量之间相关关系的本质。观察到的相关关系有可能是由潜在变量造成，因而造成误导，潜在变量一向都有这种影响。有的情况下，在潜在变量的每个等级都出现的相关关系，当我们把各等级整合在一起时却消失不见或甚至改变了方向，这就是辛普森悖论。

卡方检验可以用来判断，双向表里所出现的相关关系是否有统计显著性。卡方统计量是对“双向表中的计数”与“当列变量及行变量之间没有相关关系时我们会预期的计数”的两个计数做比较。该统计量的抽样分布不是正态分布。它是一种新的分布，叫做卡方分布。



第 24 章 习题

24.1 课外活动和成绩。北卡罗来纳州立大学检视了主修化工的学生在某必修课中的表现。他们感兴趣的问题之一，是一个学生花在课外活动的时间，和该生在该科目是否得到 C 以上的成绩，二者之间有多大的相关关系。以下是回答了有关课外活动问题的 119 位学生的相关数据：

	课外活动(每周小时数)		
	<2	2—12	>12
C 或更佳	11	68	3
D 或 F	9	23	5

将可以描述花在课外活动的时间，和修课成绩之间相关关系的百分比计算出来。用简单的叙述做个结论。

24.2 学生和其父母的吸烟习惯 学生的吸烟习惯和父母是否吸烟有关系吗？以下是对亚利桑纳州八所中学学生调查所得结果的双向表：

	吸烟学生	不吸烟学生
父母均吸烟	400	1 380
父母中有一人吸烟	416	1 823
父母均不吸烟	188	1 168

对此题中提出的问题做个简短的回答，其中要包括对某些特定百分比的比较。

24.3 蟒蛇蛋。水蟒的蛋孵不孵得出来，是否会受蛇巢所在处的温度影响？研究者把一些刚生出来的蛋分配到三种温度：暖、中等或者冷。暖这个等级相当于蟒蛇妈妈所提供的温度，冷则相当于蟒蛇妈妈不在场的情况。以下是不同温度的蛋数及孵出蛋数的资料：



	蛋数	孵出数
冷	27	16
中等	56	38
暖	104	75

- (a) 造一个温度对应结果(是否孵出)的双向表。
- (b) 计算每一组的孵出百分比。研究者认为蛋在冷水里会孵不出来, 而数据是否支持这种想法?

24.4 枪击死亡。美国的非疾病死因中, 枪击占第二位, 仅次于机动车辆。下面列有 1990—1994 年间, 对威斯康星州密尔沃基市所有与枪击相关的死亡所做研究得到的计数。我们想比较杀人和自杀事件中所用的枪支种类。我们猜想自杀会比杀人更常用长枪(猎枪及来福枪), 因为很多人在家里备有这种枪以备打猎之用。用柱状图仔细比较一下杀人和自杀的状况。长枪和手枪的使用有什么不同吗?

	手枪	猎枪	来福枪	未知	总数
杀人	468	28	15	13	524
自杀	124	22	24	5	175

24.5 谁得到学位? 不同性别之间对学位的追求有什么差别? 在下面的表里面, 列出了《美国统计精粹》当中的 1996 年所有学位资料(计数单位为千人), 资料根据不同学位以及获得学位者的性别做了分类。

	学士	硕士	专业	博士
女性	642	227	32	18
男性	522	179	45	27
总数	1 165	406	77	45

- (a) 共有多少人得到学士学位? 你觉得为什么得到学士学位的总人数并不等于得到学士学位的男性和女性的总和?
- (b) 每一种学位各有多少百分比是由女性得到? 请简短描述, 从数据可以看出性别和学位之间有何种相关关系。



24.6 商科男、女学生的主修方向。一项探讨年轻男女职业生涯规划的研究向伊利诺伊大学企管学院的所有 722 位四年级生发出了问卷。其中有个问题是问学生在企管领域主修什么。以下是填答了问卷的学生资料：

	女性	男性
会计	68	56
管理	91	40
经济	5	6
金融	61	59

用以下三种方法来描述男学生和女学生的主修方向分布的差别：百分比、图以及用话说明。

24.7 只看总数是不够的。以下两列两行的双向表中，列出了列总和及行总和：

a	b	50
c	d	50
60	40	100

找出两组不同的计数 a 、 b 、 c 、 d 放在表当中能符合所列出的列总和及行总和。这代表两个变量之间的关系，不能只从该两个变量的个别分布得到。

24.8 航班延误。以下是两家航空公司 1 个月之内在 5 个机场的航班准时及延误的计数。各航空公司的整体准时百分比常会出现在新闻报道中。机场是一项潜在变量，有可能对这类报道产生误导。

	阿拉斯加航空		美国西部航空	
	准时	延误	准时	延误
洛杉矶	497	62	694	117
凤凰城	221	12	4 840	415
圣迭戈	212	20	383	65
旧金山	503	102	320	129
西雅图	1 841	305	201	61



- (a) 阿拉斯加航空的所有航班中有多少百分比是延误的?美国西部航空的所有航班中有多少百分比是延误的?新闻中报道的通常是这些数字。
- (b) 现在算出阿拉斯加航空在每个机场的延误百分比。再算出美国西部航空的延误百分比。
- (c) 美国西部航空在每个机场的表现都比较差,但整体表现却比较好。听起来似乎不可能。仔细根据数据来解释,为何会发生这种状况。(凤凰城和西雅图的天气状况,是这个辛普森悖论例子的背后“元凶”。)

24.9 种族及死刑。被定罪的谋杀犯是否会被判死刑,似乎和被害人的种族有关。以下是被告被判谋杀的 326 件案件的相关数据:

	被告为白人			被告为黑人	
	被害人 为白人	被害人 为黑人		被害人 为白人	被害人 为黑人
死刑	19	0	死刑	11	6
非死刑	132	9	非死刑	52	97

- (a) 根据这些数据造一个被害人的种族(白人还是黑人)对应死刑(是否被判死刑)的双向表。
- (b) 验证上面的数据是否符合辛普森悖论:整体来说白人被告被判死刑的百分比较高,然而对黑人受害者和白人受害者分别来看,黑人被告被判死刑的百分比却较高。
- (c) 用数据来说明为什么辛普森悖论成立,请以法官会了解的话来解释。

24.10 商科男、女学生的主修方向。问题 24.6 列出了发给所有 722 位商学院毕业班学生的问卷结果。

- (a) 其中有两格的观测计数偏小。这些数据是否满足适合使用卡方检验的标准?
- (b) 商科学生的性别和主修方向之间,有没有具有统计显著性的相关关系。
- (c) 有多少百分比的学生没有回应这项问卷调查?无回应会削弱根据问卷结果所做的结论。



24.11 课外活动与学业成绩。你在习题 24.1 中描述了课外活动和学生在必修课表现好不好之间的相关关系。我们所观察到的这两个变量间的相关关系，是否具统计显著性？要知道答案，就照下面步骤前进。

- (a) 把习题 24.1 的双向表的列总和及行总和算出来，并算出每一格的预期计数。哪些观察到的计数距预期计数的差距最大？
- (b) 算出卡方统计量。哪几格对这个统计量的“贡献”最大？
- (c) 自由度是多少？利用表 24.1 来说明卡方检验的统计显著性，替这项研究写个简短结论。

24.12 学生和其父母的吸烟习惯。在习题 24.2 中你已看到父母和孩子的吸烟习惯之间有相关性。父母吸愈多烟，孩子愈有可能吸烟。我们想知道这项相关关系是否具统计显著性。

- (a) 把卡方检验的两项假设写出来。你觉得总体是什么？
- (b) 算出所有的预期格计数。用简单易懂的语言叙述“预期计数”什么意思。
- (c) 找出卡方统计量及其自由度。对统计显著性有何结论？

24.13 蟒蛇蛋。习题 24.3 里有在三种不同温度的水里孵蟒蛇蛋的相关数据。温度对于孵化有显著效应吗？把你的计算过程和结论做个清楚的总结。

24.14 压力和心脏病猝发。你读到报纸上的一篇文章，内容是关于调适压力是否能减少心脏病猝发的研究。全部 107 位受试者流向心脏的血流量均低于常人，所以都有心脏病猝发的风险。他们被随机指派到三个组。文章稍后谈到：

其中一组参加了为时 4 个月的压力调适计划，另一组参加了为时 4 个月的运动计划，而第三组只有从各自的医师处得到对心脏病患的一般照顾。

在接下来的 3 年内，压力调适组的 33 个人中，只有 3 个人曾经历“心脏事件”，这是指致命或非致命的心脏病猝发，或者曾接受诸如心脏搭桥术或血管扩张术。同一段期间，运动组的 34 人中有 7 人，一般照顾组的 40 人中有 12 人经历了这些事件。

- (a) 利用该篇文章中的信息，造一个描述研究结果的双向表。
- (b) 这三种处理在预防心脏病事件上，成功率各是多少？
- (c) 找出在各处理无差别原假设之下的预期格计数。请验证这些预



期计数符合我们所订使用卡方检验的规范。

(d) 这三种处理的成功率之间，有具显著性的差异吗？

24.15 儿童照护的标准。不受规范约束的家庭托儿保姆，在不同的城市是否有不同的健康安全惯例？一项研究检视了美国三个城市贫穷区域中定期受雇的保姆，其中曾需取得父母同意以便送小孩去急诊室治疗的人：在新泽西州纽瓦克市，73 个保姆中有 42 人；同州卡姆登市是 101 人中有 29 人；而伊利诺伊州的南芝加哥是 107 人中有 48 人。

(a) 用卡方检验来判断，在三个城市中需要取得父母医疗授权的保姆比例，有没有具显著性的差异？你的结论是什么？

(b) 数据要怎么产生，你的检验才能站得住脚？(事实上这项研究中有一部分样本，是询问另一项研究中的受试对象，谁替他们照顾小孩而得到的。这项研究的作者，很明智的没有做统计检验，他写道：“如果应用随机样本所适用的传统统计方法，可能会得到偏差及误导的结果。”)

第 25 章

有关总体平均数的推论*

体温 98.6°F 正常吗?

我们都听说过，把温度计放在舌下量时得到的“正常体温”应该是 98.6°F (也就是 37°C)。当然你实际的体温在一天当中会上上下下。通常早上 6 点左右体温会最高，而大约下午 4—6 点之间体温最低。不同的人之间体温也有差别，而且小孩子会比较高些。所以 98.6°F 其实是指平均温度。很可能它的意思是说，如果我们在一天当中，对所有健康成年人度量许多次体温的话，得到的平均温度会是 98.6°F 。

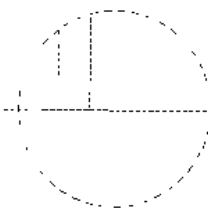
这项声明正确吗？有些很积极的医师帮 148 位健康成年人 (18—40

* 本章内容精深，要用到第 21 章中选读的部分。



岁之间)测量体温,一天量四次,共持续三天。他们声称传统的数字错了,正确的平均是 98.2°F 而不是 98.6°F 。旧的这个数据可以追溯到 1861 年,也真的是够老了。我们可能会问,新数据虽然只从 148 个人得到,但温度计比以前精确得多,能否当作提出一个新数值的充分理由。这是总体平均数的推论问题。置信区间可以回答“我们对平均数的真正值可以下怎样的结论”的问题。而显著性检验可以回答“真正的平均数并不是 98.6°F 吗”的问题。

只要我们把数据输进电脑,它就会替我们执行推论。此外我们也该问些电脑不会替我们问的问题。我们为什么对平均数感兴趣?我们真正想知道的,难道不是体温到了多高或多低,就代表身体可能出问题了吗? 98.2 和 98.6 之间的差距,是不是小到在医学上不具重要性? 1861 年的结果是不是四舍五入到整数(37°C)之后,再转换成 98.6°F ,因此它的准确程度被高估了?这些都是很好的问题,这些问题比新做的研究还要更能提醒我们,“ 98.6°F 是正常体温”说得太简略了,对医师和病人都没多大用处。



虽然置信区间和显著性检验背后的论理依据是一样的,但对于不同的特定问题,处理方法却大不相同。推论过程的形式,首先和你想从中寻求信息的参数有关,这个参数也许是总体比例或平均数或中位数或任何别的数。第二项影响因素是抽样或实验的设计。要用分层样本来估计总体比例,使用的公式就和 SRS 的不同。本章所提出对总体平均数做推论用的置信区间和显著性检验,是针对数据为抽自总体的 SRS 的情况。

样本平均数的抽样分布

你大学里的大一学生,平均一星期念几小时的书?他们在高中时的平均成绩又是多少?我们常常会要估计一个总体的平均数,为了区别总体平均(是参数)和样本平均 \bar{x} ,我们把总体平均用符号 μ (希腊



字母, 读音为 miu) 表示。我们用一个 SRS 的平均 \bar{x} , 来估计总体的未知平均数 μ 。

就同样本比例 \hat{p} 一样, 从一个较大 SRS 所得到的样本平均 \bar{x} , 其抽样分布会接近正态分布。因为 SRS 的样本平均是 μ 的无偏估计量, 所以 \bar{x} 的抽样分布的平均数就是 μ 。 \bar{x} 的标准差由总体的标准差决定, 后者通常用符号 σ (希腊字母, 读音为 sigma) 表示。我们可以用数学得出下列事实。

• 样本平均数的抽样分布

从平均数为 μ , 标准差为 σ 的总体抽取大小为 n 的 SRS。用 \bar{x} 表示样本平均数。则:

- 当样本大小 n 较大时, \bar{x} 的抽样分布会近似正态分布。
- 抽样分布的平均数等于 μ 。
- 抽样分布的标准差是 σ/\sqrt{n} 。

许多样本所算出的 \bar{x} 值, 都是以总体平均 μ 为中心, 这应该不令人意外。这只是随机样本无偏性质的再度表现。抽样分布的另两件事实, 则具体展现了样本平均 \bar{x} 的两项极重要性质:

- 一些观测值的平均, 比个别观测值的变化要小。
- 一些观测值的平均的分布, 要比个别观测值的分布更接近正态。

图 25.1 描述了上述的第一个性质。它比较了 1 个观测值的分布和 10 个观测值平均数 \bar{x} 的分布。两者的中心位置相同, 但是平均数 \bar{x} 的分布比较集中。在图 25.1 里, 个别观测值的分布是正态分布。在这种状况之下, 不管样本的大小如何, \bar{x} 的抽样分布都是确实的正态分布, 而不只是在样本大时为近似正态分布。有一项了不起的统计事实, 叫做**中央极限定理**(central limit theorem), 说的是如果我们从任何总体随机抽取愈来愈多的观测值, 则这些观测值平均数的分布, 迟早会接近正态分布。(这项伟大的事实要成立, 必须符合某些技术条件, 不过我们在应用时可以将之略去不管。)用正态分布当作样本平均的抽样分布, 所依据就的是中央极限定理。

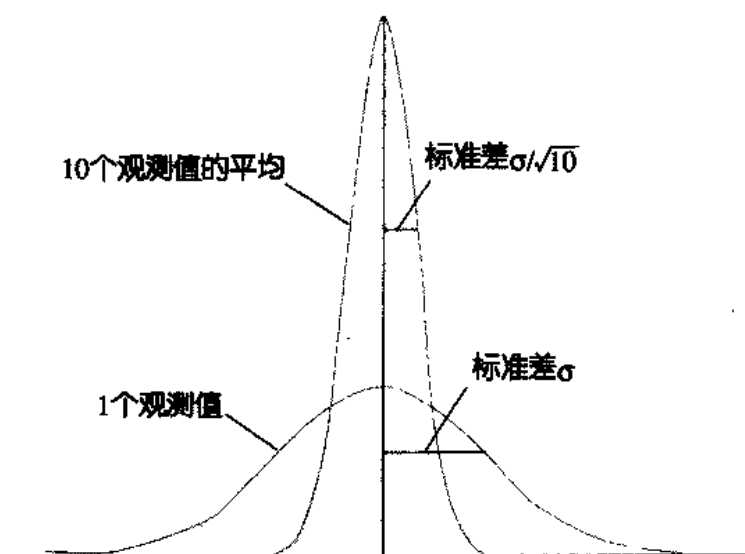


图 25.1 10 个观测值的平均数 \bar{x} 的抽样分布和个别观测值的分布的比较

例 1 中央极限定理的行动

图 25.2 呈现了中央极限定理的“作用”过程。左上角的密度曲线描绘的是抽自某个总体的个别观测值的分布。该分布为强烈右偏。举例来说，这类分布可用来描述修理某项家电所需要的时间。其中大部分都可以很快修理好，但有一些要花很多时间。

图 25.2 里的另外三条密度曲线，则分别代表从同一总体抽出 2 个、10 个及 25 个观测值所得样本平均的抽样分布。随着样本大小 n 增加时，密度曲线的形状会愈来愈接近常态，其平均数保持不变，但标准差会遵循 σ/\sqrt{n} 规则而愈来愈小。10 个观测值的分布还是有点右偏，但已开始像正态曲线了，而 $n=25$ 的密度曲线就更接近正态。总体分布的形状*和 10 或 25 个观测值平均数的分布形状，其间的对比极其明显。

*译注：即 $n=1$ 时的分布。

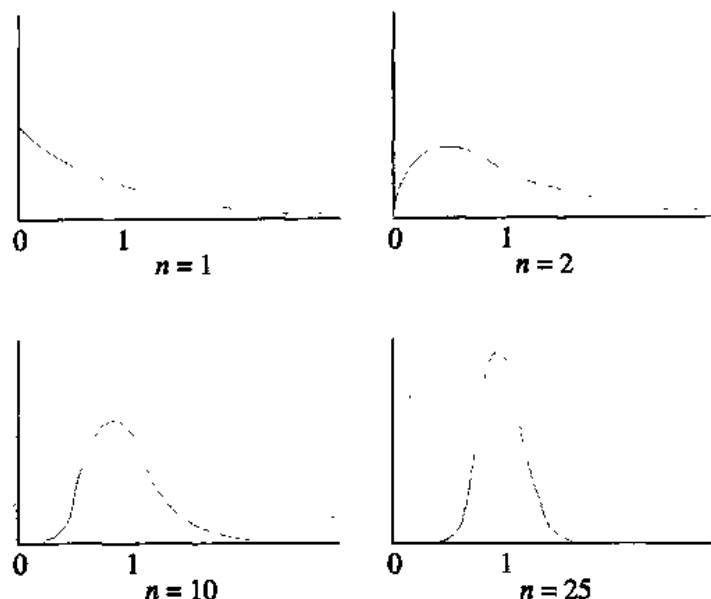


图 25.2 当样本大小增加时, 样本平均数 \bar{x} 的分布会愈来愈接近正态分布。个别观测值的分布 ($n=1$) 完全不是正态分布。样本平均数的分布随着观测值的个数, 从 2 个、10 个到最后增加为 25 个, 而愈来愈接近正态分布

总体平均数的置信区间

\bar{x} 的标准差决定于样本大小 n 以及总体中个体的标准差。我们知道 n 多大, 可是不知道 σ 是多少。当 n 很大的时候, 样本标准差 s 会接近于 σ , 因此可以用来估计 σ , 就像我们用样本平均 \bar{x} 来估计总体平均 μ 一样。因此 \bar{x} 的估计标准差就是 s/\sqrt{n} 。现在我们可以照着第 21 章中导出比例 p 的置信区间一样的推理过程, 找出 μ 的置信区间。主要是为了要涵盖正态曲线底面积为 C 的中间部分, 我们必须从平均数往两边各延伸 z^* 的距离。再看一下图 21.5 里, C 和 z^* 之间是什么关系。

总体平均数的置信区间

从个体平均数为 μ 的大总体里, 抽取大小为 n 的 SRS。样本中观测值的平均是 \bar{x} 。当 n 很大的时候, μ 的近似水平 C 置信区间为

$$\bar{x} \pm z^* \frac{s}{\sqrt{n}}$$

此处 z^* 为对应置信水平 C 的临界值, 是从表 21.1 得到的。



在估计 p 时提过的注意事项, 在这里也适用。只有在样本是 SRS 而且样本大小 n 够大时, 才可以用这个公式。当样本大小 n 增加时, 误差界限减小的比例, 还是只跟 \sqrt{n} 成比例。再提醒一件事: 要记住 \bar{x} 和 s 都会受异常值严重影响。当异常值存在时, 用 \bar{x} 和 s 做的推论就有疑问。永远要检视你的数据。

例 2 全国教育成果评估的数学分数

全国教育成果评估 (NAEP) 中有一项对掌握数字技巧的简短测验, 内容主要是基本算术和把算术应用到实际问题的能力。测验的可能分数从 0 分到 500 分。举例来说, 得 233 分的人会算得出银行存款单上两张支票金额的和; 考 325 分的人能根据菜单算出吃一餐要花多少钱; 考 375 分的人, 会把价钱从每盎司要价几美分转换成每磅几美元。

最近有一年的 NAEP 样本包括了 840 位 21—25 岁的男性。他们的平均数学分数是 $\bar{x} = 272$, 标准差是 $s = 59$ 。这 840 位男性是从总体为所有年轻男性得来的一个 SRS。根据这个样本的结果, 我们可以对这个年龄层全部 950 万名男性的总体平均分数 μ 做怎样的结论?

μ 的 95% 置信区间所用的临界值是从表 21.1 里得到的 $z^* = 1.96$ 。信赖区间为:

$$\begin{aligned}\bar{x} \pm z^* \frac{s}{\sqrt{n}} &= 272 \pm 1.96 \frac{59}{\sqrt{840}} \\ &= 272 \pm (1.96)(2.036) = 272 \pm 4.0\end{aligned}$$

我们有 95% 信心, 所有年轻男性的平均分数在 268—276 之间。

抓作弊

许多学生有机会考到题目很多的选择題測驗。电脑阅卷时有没有可能一边评分, 一边揪出答案接近到令人怀疑的考卷呢? 有些很聪明的人已发明了评估的方法, 不仅考虑到答案是否相同, 也考虑到各个答案会有多少人选择, 以及相似考卷的总得分。这项评估接近正态分布, 两张考卷的评估如果超过 ± 4 个标准差, 就会被电脑记录为有统计显著性。



总体平均的检验

就像置信区间一样，对有关总体平均数 μ 的假设所做的显著性检验，背后的理论依据和有关总体比例 p 的检验相同。主要的概念是要用到当原假设为真时，样本平均 \bar{x} 的抽样分布。找出从你样本得到的 \bar{x} 值在这个分布的位置，看看是否不容易发生。 H_0 为真时很难得出现的 \bar{x} 值，就是 H_0 不正确的根据。检验的四个步骤也和检验比例时类似。以下有两个例子，第一个是单边检验，第二个是双边检验。

例 3 你会不会算支票簿的存款余额？

在讨论美国劳动人口的教育程度时，一位悲观者说：“普通的年轻人，连计算支票簿的存款余额都不会。”全国教育成果评估 (NAEP) 的调查结果声称，在他们的数学测验中得到 275 分以上的人，就具备计算支票存款簿余额所需的技巧。NAEP 随机样本的 840 位年轻人得到的平均分数是 $\bar{x} = 272$ ，比计算存款余额所需的标准稍低一些。这项样本结果是否足以当做所有年轻人的平均都低于 275 的证据？样本中分数的标准差为 $s = 59$ 。

假设。悲观者的断言是，NAEP 的平均分数低于 275。这是我们的备择假设，我们要找证据来支持这个说法。两项假设分别是：

$$H_0 : \mu = 275$$

$$H_a : \mu < 275$$

抽样分布。如果原假设为真，样本平均数 \bar{x} 会接近平均数为 $\mu = 275$ 、标准差为：

$$\frac{s}{\sqrt{n}} = \frac{59}{\sqrt{840}} = 2.036$$

的正态分布。这次我们又是用样本标准差 s 来代替未知的总体标准差 σ 。

数据。NAEP 样本的平均是 $\bar{x} = 272$ 。这项结果的标准计分为：

$$\begin{aligned}\text{标准计分} &= \frac{\text{观测值} - \text{平均值}}{\text{标准差}} \\ &= \frac{272 - 275}{2.036} = -1.47\end{aligned}$$

这是说，样本结果和我们所预期的(即平均来说，年轻人的数学能力恰恰只够计算支票簿的存款余额)，大约是低了 1.47 个标准差。

P 值。图 25.3 把样本结果 -1.47(以标准计分为刻度)标示在代表 H_0 为真时的抽样分布的正态曲线上。这条曲线用的是标准刻度(standard scale)，所以平均数为 0，标准差为 1。我们单边检验的 P 值，是 -1.47 左边阴影区的面积。为了查表 B，我们把标准计分四舍五入到 -1.5。表 B 表里面说，-1.5 是 6.68 百分位数，所以它左边的面积就是 0.0668。这就是我们的 P 值。(这个结果是近似值，因为我们为了要使用表 B，已把标准计分四舍五入。而用电脑软件可得 $P=0.071$ 。)

结论。大约 $P=0.07$ 的 P 值显示，似乎所有年轻人的平均分数，并未达到计算存款余额的标准，但是证据不能说很强。

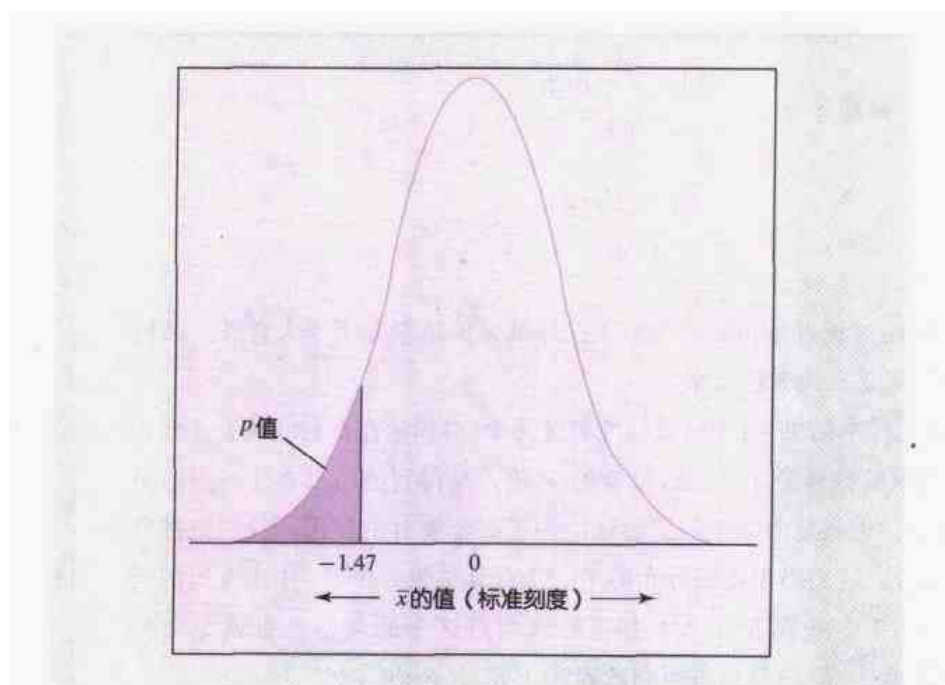


图 25.3 当样本平均数的标准计分为 -1.47 时，单边检验的 P 值



例 4 主管的血压

国家健康统计中心提出报告说, 35—44 岁男性的平均收缩血压是 128。某大公司的保健主任检视了上述年龄层的 72 位主管的病历记录, 发现这个样本的平均收缩血压是 $\bar{x} = 126.1$, 而标准差是 $s = 15.2$ 。这可不可以当做该公司主管的平均血压和一般大众不同的证据?

假设。原假设是和全国平均“没有差别”。备择假设是双边的, 因为该保健主任在检视数据之前, 心里并没有特定的方向。所以有关所有主管的总体的未知平均数 μ 的假设为

$$H_0: \mu = 128$$

$$H_A: \mu \neq 128$$

抽样分布。如果原假设为真, 样本平均数 \bar{x} 就会近似正态分布, 其平均数为 $\mu = 128$, 标准差为

$$\frac{s}{\sqrt{n}} = \frac{15.2}{\sqrt{72}} = 1.79$$

数据。样本平均为 $\bar{x} = 126.1$, 这个结果的标准计分是

$$\begin{aligned} \text{标准计分} &= \frac{\text{观测值} - \text{平均值}}{\text{标准差}} \\ &= \frac{126.1 - 128}{1.79} = -1.06 \end{aligned}$$

我们知道, 只刚超过正态分布的平均数 1 个标准差的结果并不令人意外。最后一个步骤只是把这个观念正式表达出来。

P 值。图 25.4 把样本结果 -1.06 (以标准刻度为单位) 标示在正态曲线上, 此正态曲线代表 H_0 为真时的抽样分布。双边检验的 P 值, 是得到的结果在任一方向至少达到这么远的概率, 也就是曲线下的阴影区。为了配合表 B 的使用, 我们把标准计分四舍五入到 -1.1。这相当于正态分布的 13.57 百分位数。所以 -1.1 左边的面积是 0.1357。在 -1.1 左边以及在 1.1 右边的面积是这个的两倍, 也就是大约 0.27。这是我们的近似 P 值。(从软件得到的确实 P 值是 $P = 0.289$ 。)

结论: 这么大的 P 值让我们没有理由认为, 所有主管的平均血压和全国同年龄层的平均血压有所不同。



这项检验是假设样本中的 72 位主管，是得自该公司所有中年男性主管总体的一个 SRS。我们应该问问数据怎么来的，以便检验这项假设是否成立。比如说，如果只有最近身体出过问题的主管才有病历记录可查，这些数据对我们要检验的问题就没什么用处。结果发现是每位主管每年都可做一次免费的健康检查，而保健主任是从健康检查结果中随机选出了 72 件。

例 4 中的数据并没有说，这家公司主管的平均血压是 128，我们试图寻找 μ 不等于 128 的证据，但是没找到足以令人信服的证据。因此我们只能够下这样的结论。整个主管总体的平均血压，无疑的必定不会恰好等于 128。样本只要够大，一定会出现不相等的证据，即使差距很小。

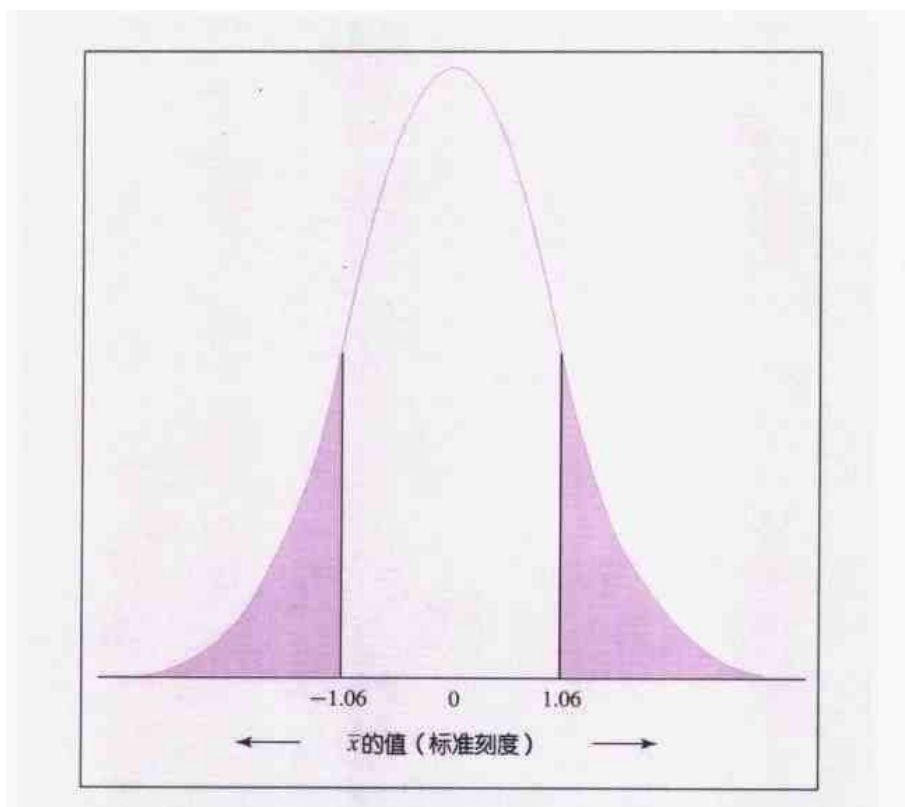


图 25.4 当样本平均的标准计分为 -1.06 时，双边检验的 P 值



本章重点摘要

我们用取自总体的 SRS 的样本平均 \bar{x} ，来估计总体平均 μ 。 μ 的置信区间和显著性检验，是根据 \bar{x} 的抽样分布得来的。当样本大小 n 较大时，**中央极限定理**告诉我们这个抽样分布大致是正态分布。虽然方法有细节上的差异，但是关于 μ 的推论方式，很像关于总体比例 p 的推论，因为二者都是根据正态抽样分布而来的。



第 25 章 习题

25.1 抽样分布的概念。图 21.1 用图表示了样本比例 \hat{p} 的抽样分布的概念。画一个类似的图，来呈现样本平均 \bar{x} 的抽样分布的概念。

25.2 主管的血压。例 4 里面算出，一个样本中的主管的平均血压，和全国平均血压 128 之间，没有达到 10% 的统计显著性水平的差异。用该例子中的数据造一个主管的总体平均血压的 90% 置信区间。为什么你会预期 128 将落在这个区间里面？

25.3 IQ 测验分数。以下是美国中西部一个学区中，31 个七年级女学生的 IQ 测验分数：

114	100	104	89	102	91	114	114	103	105	
108	130	120	132	111	128	118	119	86	72	
111	103	74	112	107	103	98	96	112	112	93

- 我们认为 IQ 分散的分布会接近正态分布。画出这 31 个分数分布的直方图。你的图有没有显示出异常值、明显的偏斜或其他不符合正态分布的特点？用计算机算出这些分数的平均数和标准差。
- 把这 31 个女生当做该学区中所有中学女生的一个 SRS，选一个总体平均分数的 95% 置信区间。
- 事实上这些分数是该学区当中某一所学校内，所有七年级女生的分数。详细说明为什么 (b) 部分算出的置信区间不可靠。

25.4 置信水平及误差界限。NAEP 测验(例 2)也有包括 1 077 位 21—25 岁女性的样本。她们的平均分数是 275，标准差是 58。

- 选一个所有年轻女性平均分数 μ 的 95% 置信区间。
- 算出 μ 的 90% 和 99% 置信区间。
- 分别对应置信水平 90%、95% 和 99% 的误差界限是多少？把置信水平提高，会对置信区间的误差界限产生怎样的影响？

25.5 IQ 测验分数。任何年龄层的整个总体的平均 IQ 分数都应该



是 100。把习题 25.3 里的 IQ 分数当做来自该学区所有中学女生的一个 SRS。这些分数是否提供足够证据, 显示总体的平均 IQ 不是 100?

25.6 平均和个别值的对照。美国大学测验 (ACT) 这项大学入学考试的分数, 变化情况符合平均数 $\mu = 18$ 、标准差 $\sigma = 6$ 的正态分布。已知分数的范围是从 1 到 36。

- (a) 个别分数的中间 95%, 会在怎样的范围?
- (b) 如果我们把随机选出的 25 个学生的 ACT 分数平均, 则平均分数 \bar{x} 的中间 95% 会在怎样的范围?

25.7 学生的态度、读书习惯及态度调查 (SSHA, Survey of Study Habits and Attitudes) 是一项心理测验, 用来度量学生的读书习惯和对学校的态度, 分数从 0 到 200。美国大学生的平均分数大约是 115, 标准差大约是 30。有位老师猜测, 年纪较大的学生对学校的态度会比较好。她让 25 位 30 岁以上的学生做了 SSHA 测验。假设较年长学生的总体的分数, 符合标准差 $\sigma = 6$ 的正态分布。这位老师想要检验以下假设:

$$H_0: \mu = 115$$

$$H_a: \mu > 115$$

- (a) 如果原假设为真, 25 位较年长学生的样本的平均分数 \bar{x} , 会有怎样的抽样分布? 把这个分布的密度曲线画出来。(提示: 先画一条正态曲线, 再根据你对于 μ 和 σ 在正态曲线上位置的了解, 在 \bar{x} 轴上做出标示。)
- (b) 假设从样本数据得到 $\bar{x} = 118.6$ 。把这个点标示在你画的图的 \bar{x} 轴上。事实上所得的结果是 $\bar{x} = 125.7$, 也把这点标示在你的图上。根据你的图用简单易懂的语言解释, 为什么其中一个结果是很好的证据, 可显示所有较年长学生的平均分数超过 115, 而另一个结果就不是。
- (c) 把曲线下代表样本结果 $\bar{x} = 118.6$ 的 P 值区域以阴影表示。

25.8 血压。有一项随机化比较实验研究了食物对血压的影响。研究者把 54 位健康白种男性随机分成两组。其中一组服用钙补充剂, 另一组服用安慰剂。在研究开始的时候, 研究人员对受试者度量了许多变量的值。在提出研究结果的论文中, 报告了安慰剂组 27 人静止



时收缩压的平均为 $\bar{x} = 114.9$, 标准差为 $s = 9.3$ 。

(a) 替受试者所属总体的平均血压造一个 95% 置信区间。

(b) 你在(a)中所用的公式, 要求数据来源的 27 人必须满足某个重要假设。这项假设是什么?

25.9 迷宫中的小鼠。有关动物研究的实验当中, 有一种是度量小鼠要花多少时间才能走出迷宫。对某一特定迷宫来说, 平均所需时间是 18 秒。有一位研究者认为, 如果加上很大的噪音, 应该会让小鼠更快走出迷宫。她度量了在噪音刺激之下, 好几只小鼠分别完成迷宫所需的时间。原假设 H_0 和备择假设 H_a 分别是什么?

25.10 回应时间。去年你公司的维修技术员, 对签了维修合同的顾客报修之后的平均回应时间是 2.6 小时。今年的数据是否显示出平均回应时间有具统计显著性的改变? 要回答这个问题, 你应该检验怎样的原假设和备择假设?

25.11 检验随机数字产生器。我的统计软件中有“随机数字产生器”, 它产生的数字照理应该是在 0—1 之间随机散布。如果这属实, 则生产出来的数字可视为来自 $\mu = 0.5$ 的总体。在给予产生 100 个随机数字的指令之后, 所得到的结果平均数字为 $\bar{x} = 0.532$, 标准差为 $s = 0.316$ 。这够不够当做证据, 显示该软件产生的所有数字的平均并非 0.5?

25.12 他们会不会刷更多? 某银行想要知道, 如果对一年至少刷 2 400 美元的信用卡顾客免收年费, 是否会增加他们的刷卡金额。该银行从信用卡客户中抽出 200 人的 SRS, 提供这项优惠。然后银行比较了这些客户今年和去年的刷卡金额。样本增加的刷卡金额平均为 332 美元, 标准差 108 美元。对于免年费会增加平均刷卡金额, 这些数据是否提供了有 1% 的统计显著性水平的证据? 写出 H_0 和 H_a 并做检验。

25.13 检验随机数字产生器。替习题 25.11 的软件所生产的所有随机数字的平均造一个 90% 置信区间。

25.14 他们会不会刷更多? 习题 25.12 报告了某银行对至少刷 2 400



美元的信用卡客户免年费的试验结果。如果这项优惠实验遍及所有客户时，为平均会增加的刷卡金额造一个 99% 置信区间。

25.15 抽样分布、习题 25.11 和 25.13 考虑了用电脑程序所产生的随机数字的平均。因为这些数字应该在 0—1 之间随机散布，所以它们的平均应该是 0.5。我下指令叫软件不断生产出包含 100 个随机数字的样本。以下是 50 个大小为 100 的样本平均 \bar{x} 的值：

0.532 0.450 0.481 0.508 0.510 0.530 0.499 0.461 0.543 0.490
0.497 0.552 0.473 0.425 0.449 0.507 0.472 0.438 0.527 0.536
0.492 0.484 0.498 0.536 0.492 0.483 0.529 0.490 0.548 0.439
0.473 0.516 0.534 0.540 0.525 0.540 0.464 0.507 0.483 0.436
0.497 0.493 0.458 0.527 0.458 0.510 0.498 0.480 0.479 0.499

\bar{x} 的抽样分布，是指所有可能样本的平均数的分布。我们有 50 个样本的平均数，画出这 50 个观测值的直方图。分布看起来是否大致像正态，就如中央极限定理所说样本够大时应该发生的情况？

25.16 他们会不会刷更多？在习题 25.12 和 25.14 当中，你替某银行进行改变信用卡规则的实验，做了检验并算出置信区间。你对这项研究应该要问一些问题。

- (a) 刷卡金额的分布是右偏的，但是不会有异常值，因为银行对每张卡的额度都设有限制。为什么我们还可以根据样本平均数 \bar{x} 的正态抽样分布来做检验和计算置信区间呢？
- (b) 银行做的实验并不是比较实验。今年刷卡金额多于去年，也许可以用潜在变量来解释，而不是由于规则的改变。刷卡金额增加有哪些可能理由？简单描述可以回答银行问题的随机化比较实验的设计。

25.17 抽样分布，续集。习题 25.15 里有 50 个大小为 100 的随机样本的 50 个样本平均数 \bar{x} 。用计算机算出这 50 个数的平均及标准差，然后回答下列问题。

- (a) 如果随机数字产生器正确的话，这 50 个样本所来自的总体，其平均应该是 $\mu = 0.5$ 。你预期所有可能样本的 \bar{x} 的分布，其平均数会是多少？这 50 个样本的平均数距离这个数近吗？



- (b) 样本大小为 $n = 100$ 的分布，其标准差应该是 $\sigma/10$ ，而此处的 σ 是总体中个体的标准差。用这事实以及你从 50 个 \bar{x} 所算出的标准差来估计 σ 的值。

第四部分 复习

统计推论是根据样本数据来对总体做结论，并利用概率来表示结论的可靠程度。置信区间用来估计未知参数。显著性检验则告诉我们，针对某个参数的断言，证据有多强。第 21 和 22 章说明了置信区间和检验的理论依据，也提出了对总体比例 p 做推论的细节，以供需要的读者做参考。

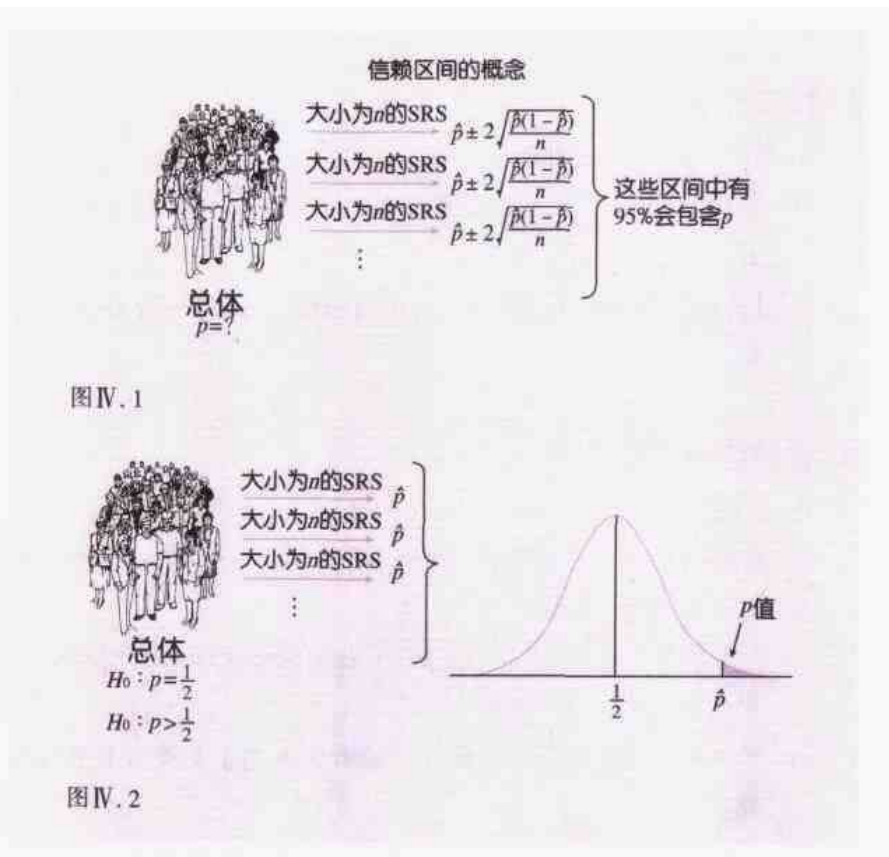
置信区间和检验里面用到的概率是告诉我们，如果我们重复使用置信区间或者检验的公式许多次，所会发生的情况。置信水平是指：用置信区间的公式所产生的区间，会抓到未知参数的概率。95% 置信区间是说，如果我们重复用这个公式，则有 95% 的时候会得到正确结果*。图 IV.1 显示出总体比例 p 的近似 95% 置信区间是什么意思。

* 译注：即得到参数。

一项检验在原假设为真时，会产生像我们观测到的结果这样极端或更极端的概率，就是我们所说的 P 值。图 IV.2 就在说明这个概



念, 把根据我们样本得到的样本比例 p , 标示在当原假设成立时, 显示所有样本变化情况的正态曲线上面。 P 值告诉我们观测到的结果令人意外的程度。而很叫人意外的结果(小的 P 值)就是原假设不正确的合理证据。



要分得出推论的用法正确或不正确, 你必须懂得背后的道理, 还要注意一些细节以及陷阱。第 23 章就是教你这些。在选读的第 24 和 25 章当中, 另外讨论了几个特定的推论问题。第 24 章讲的是双向表, 有描述的部分也有推论的部分。此章的描述部分是在讲类别变量之间的关系, 把第二部对变量之间关系的讨论补充完整。第 25 章讲的是有关总体平均数的推论。



第四部分 重点摘要

以下是你读完 21—25 章后，应该要有的最重要技能。

A. 抽样分布

1. 能说明抽样分布的概念。参考图 21.1。
2. 会用样本比例 \hat{p} 的正态抽样分布以及 68–95–99.7 规则找出和 \hat{p} 相关的概率。
3. 【选读】会用样本平均数 \bar{x} 的正态抽样分布找出和 \bar{x} 相关的概率。

B. 置信区间

1. 能说明置信区间的概念。参考图 IV.1。
2. 能够用非专业术语说明，统计报告中的“95% 信心”及其他置信叙述所代表的意义。
3. 会用基本公式 $\hat{p} \pm 2\sqrt{\hat{p}(1-\hat{p})/n}$ 找出总体比例 p 的近似 95% 置信区间。
4. 了解置信区间的误差界限如何随着样本大小及置信水平而改变。
5. 会察觉得出应用推论时的一些错误做法，比如数据产生方法不恰当、从许多结果中选择最好的、不回应率很高却置之不理或者忽略异常值。
6. 【选读】会用较一般的公式 $\hat{p} \pm z^*\sqrt{\hat{p}(1-\hat{p})/n}$ 以及正态分布的临界值 z^* ，找出总体比例为 p 的置信区间。
7. 【选读】会用 $\bar{x} \pm z^*s/\sqrt{n}$ 这个公式找出总体平均数 μ 的置信区间。

C. 显著性检验

1. 会说明显著性检验的概念。参考图 IV.2。
2. 当我们要检验的参数是总体比例为 p 时，会写出原假设及备择假设。
3. 已经知道一项检验的 P 值时，会用非专业语言说明 P 值的意义。



4. 会说明“有 5% 的统计显著性水平”以及其他类似有关显著性的叙述的意义。会说明为什么于某个特定水平比如 5% 的统计显著性水平，所提供的信息不如 P 值多。
5. 了解显著性检验并不能度量一项效应的大小或重要性。
6. 了解并能说明小样本和大样本对于结果的显著性的影响。
7. 【选读】会用表 B 中正态分布的百分位数，算出对于比例 \bar{p} 的检验的 \bar{p} 值。
8. 【选读】会用样本平均 \bar{x} 以及表 B 对总体平均 μ 做单边以及双边检验。

D. 双向表【选读】

1. 会把一些个体的两个类别变量的资料，用双向表的计数表示出来。
2. 会利用双向表中的计数算出一些百分比，来描述两个类别变量间的关系。
3. 对于特定的双向表，能说明卡方检验所要检验的原假设是什么。
4. 会根据双向表算出预期格计数、卡方统计量以及自由度。
5. 会用表 24.1 的卡方分布来评估统计显著性。以及针对特定的双向表，解释检验结果。



第四部分 复习习题

复习习题都简短易答，目的是要加强你在书里学到的基本概念和技巧。

IV.1 重大事件。一个 489 位成人的 SRS 中，有 347 人从一堆选择当中选出第二次世界大战为 20 世纪最重要事件。替所有成人中认为第二次世界大战为世纪最重要事件的比例，造一个 95% 置信区间。

IV.2 酗酒问题。1 039 位成人的 SRS 中，有 374 人承认家庭中会有人有酗酒问题。替所有成人当中其家庭曾有酗酒相关问题的比例，造一个 95% 置信区间。

IV.3 重大事件。习题 IV.1 里有 489 位成人构成的一个 SRS。假设（做调查的人不知道这点）实际上所有成人当中有 70% 会选择第二次世界大战为 20 世纪最重大的事件。想像一下我们从这个总体抽取许多大小为 489 的 SRS，并记录每一个样本中选择第二次世界大战的百分比。所有这些百分比的中间 95% 的值，会在什么范围？

IV.4 酗酒问题。习题 IV.2 里有 1 039 位成人的 SRS。假设在包含所有成人的总体当中，恰好有 35% 的人会说家庭里曾发生酗酒问题。想像我们抽很多个大小为 1 039 的 SRS，对每一个样本都记录下家中曾有酗酒问题的人之所占比例 \hat{p} 。

(a) 描述我们的样本比例 \hat{p} 会有些什么样的值的抽样分布是什么？

(b) 利用这个分布以及 68-95-99.7 规则，找出所有样本当中，有超过 36.5% 的人回应说家里曾发生酗酒问题所占的近似百分比。

IV.5 喝太多了。包含 684 位喝酒的成人的 SRS 中，有 164 人承认他们“有时喝了过量的含酒精饮料。”替所有喝酒的人当中，有时会过量的比例造一个 95% 置信区间。

IV.6 轮盘。轮盘沿圆周的 38 格当中有 18 格是红色的。你观察轮



盘转许多次，并记录红色发生的次数。现在你想用这些数据来检验红色出现的概率 p 是否符合一个公平轮盘该有的值。写出你要检验的假设 H_0 和 H_a 。

IV.7 为什么不可以？表 11.1 记录了美国 50 个州中每一州 65 岁以上居民所占百分比。你可以检视一下，有 11 个州的百分比达到 14% 或更高。所以居民之中至少有 14% 为高龄者州的样本比例是 $\hat{p} = 11/50 = 0.22$ 。请说明为什么再去算总体比例 p 的 95% 置信区间没有意义。

IV.8 帮助领救济金的妈妈。有一项研究比较了两组有年幼孩子且两年前在领救济金的美国妈妈们。其中一组自愿参加了训练课程，是在当地一所职业学校举办的免费课程，并在当地的媒体做广告。另一组并未参加该训练课程。该研究发现，两组的母亲仍然在领救济金的比例有显著差异 ($P < 0.01$)。差异不仅具显著性，而且很大。报告说未参加训练课程者仍在领救济金的比例，比参加者高 $21\% \pm 4\%$ ，置信水平是 95%。假设你是美国某议员的助理，该议员关心领救济金的妈妈的处境，并向你问到这篇报告。

- (a) 用简单易懂的语言说明，有“显著差异 ($P < 0.01$)”是什么意思。
- (b) 扼要并清楚说明“95% 置信水平”是什么意思。
- (c) 研究结果并不能合理证明，要求领救济金的妈妈参加职业训练，能够大幅度降低要继续靠救济金过活的比例。向议员证明原委。

IV.9 打败医疗体系。有些医师觉得医疗体系规定限制太多，让他们无法有效地治疗病人，因此他们就从宽解释，帮助病人取得医疗体系给付。以下是从一项有关这个问题的研究中节录的一句话：“同意‘如今为了提供高品质的医疗，必须和医疗体系打赌’的医师比起不同意的医师，更多的人承认曾在给付上面做手脚 (64.3% 对 35.7%； $P < 0.01$)。”

- (a) 试向医师说明，在这项研究的背景下， $P < 0.01$ 是什么意思。
- (b) 有统计显著性的结果，仍然可能因差异太小而在实际上不重要。你怎么知道这里的结果不属于这种情况？

IV.10 另类疗法。一项美国的全国性随机调查，问了 1 500 位成人对于另类疗法的看法，这些疗法包括针灸、按摩和使用草药。其中有



660 人回答, 在传统疗法得不到应有的效果时, 愿意尝试另类疗法。

- (a) 替所有成人中愿意使用另类疗法的比例造一个 95% 置信区间。
- (b) 根据调查结果写一段新闻报道。

IV. 11 什么时候打给你比较好? 你大概也猜得到, 用电话进行的抽样调查在晚上打所得到的回应率, 比工作日白天打的要高。有一项研究在工作日的上午拨了 2 304 个随机选择的电话号码, 其中有 1 313 个有人接, 而当中只有 207 个访问成功。而在工作日晚间所拨打的 2 454 个电话, 有 1 840 个有人接, 其中 712 个达到访问目的。替工作日上午和工作日晚间有人接电话的比例分别造出 95% 置信区间。你是否有把握, 晚间的比例比较高?

IV. 12 另类疗法、习题 IV. 10 中的调查结果是否为合理证据, 显示超过 $1/3$ 的成人在传统疗法达不到预期效果时愿意使用另类疗法?

- (a) 写出要检验的两项假设。
- (b) 若你列出的原假设为真, 样本比例 \hat{p} 的抽样分布为何? 把这个分布画出来。
- (c) 把 \hat{p} 的实际值标示在曲线上。这个结果令人意外的程度, 是否足够提供不利于原假设的合理证据?

IV. 13 什么时候打给你比较好? 假设我们知道, 电话访问调查在工作日上午打的所有电话中, 有 57% 有人接。我们在工作日的晚间拨打了 2 454 个随机选择的号码, 其中有 1 840 个有人接。这是否为合理证据, 显示晚间打的电话有人接的比例较高?

- (a) 写出要检验的两项假设。
- (b) 如果你的原假设成立, 样本比例 \hat{p} 的抽样分布为何? 画出这个分布。
- (c) 把 \hat{p} 的实际值标示在曲线上。这个结果令人意外的程度, 是否足够成为不利于原假设的合理证据?

IV. 14 没有统计显著性。习题 IV. 9 中提到的那项研究, 检视了一些对于可能影响医师遵照医疗体系规则的因素。怕会被起诉的医师, 也许较少不照规则办事。研究报告说: “值得注意的是, 即使担心会因诈欺被起诉, 但仍没有影响医师使用这些伎俩 ($P=0.34$)。” 说明一下为什么 $P=0.34$ 这样的结果, 可以支持“担心被起诉并不会影



响行为”这样的结论。

IV.15 上教堂。民意调查显示，美国人中有约 40% 的人说他们上周有去做礼拜。这项结果十年来都没什么变动。但研究人们实际的行为，而不是他们自己说有没有去，却发现实际去教堂的比例远低于上述百分比。有一项研究算出了两个 95% 置信区间，一是根据天主教徒样本所说的，另一是根据实际行为的样本。以芝加哥为例，从民意调查样本得到的 95% 置信区间说，45.7%—51.3% 的天主教徒每周参加弥撒。从实际计数得到的 95% 置信区间，却说每周做弥撒的教徒在 25.7%—28.9% 之间。

- (a) 为什么我们可以预期，调查是否去教堂的结果，应会向高估真正参加比例的方向偏？
- (b) 芝加哥的调查显示有 48.5% 的天主教徒自称每周都做弥撒。为什么我们不直接说：“芝加哥的天主教徒中有 48.5% 声称每周做弥撒”，而要给个 45.7%—51.3% 的区间？
- (c) 自称和观察行为得到的结果差别很大。在上面提出的两个区间分别说“我们有 95% 信心”，是什么意思？

以下习题和 21 章及 22 章中选读的各节有关。

IV.16 重大事件。489 位成人的 SRS 中，有 347 位从列了许多事件的清单中，选择了第二次世界大战为 20 世纪最重要的事件。替所有成人中认为第二次世界大战为 20 世纪最重大事件者的比例，分别造出 90% 及 99% 置信区间。请简单解释，通过比较这两个区间和习题 IV.1 中的 95% 置信区间，可以说明哪些有关置信区间的重要事实？

IV.17 酗酒问题。一个 1039 位美国成人的 SRS 中，有 374 人说家里曾有酗酒问题。这是否可当作超过 $1/3$ 的美国成年人家里曾有酗酒问题的合理证据？清楚写出检验的五个步骤（假设、抽样分布、数据、 P 值、结论）。

IV.18 喝太多了。684 位喝酒成人的 SRS 中，有 164 人承认他们“有时喝酒过量”。替所有喝酒者中有时喝过量的比例造一个 90% 置信区间。90% 置信区间有哪点比不上 95% 置信区间？又有哪一点比 95% 置信区间好？



IV.19 对选民做抽样调查。你是某美国国会议员的民意调查顾问。一个 500 位登记选民的 SRS 中,有 28% 说“环境问题”是美国面对的最重要议题。替所有选民中有这种想法的比例造一个 90% 置信区间。然后向国会议员详细解释,对于选民的意見你的结论是什么。

IV.20 另类疗法。完成习题 IV.12 中要做的显著性检验的细节部分。要清楚写出检验的五个步骤(假设、抽样分布、数据、 P 值、结论)。

IV.21 什么时候打给你比较好?完成习题 IV.13 中要做的显著性检验的细节部分。要清楚写出检验的五个步骤(假设、抽样分布、数据、 P 值、结论)。

以下习题和选读的第 24 及 25 章有关。

IV.22 总裁的薪水。一项对 104 家公司的研究,发现公司付给总裁的薪水,实际上每年平均增加了 $\bar{x} = 6.9\%$ 。所增加百分比的标准差为 $s = 17.4\%$ 。

- (a) 104 项个别增加的百分比,分布是右偏的。说明为什么根据中央极限定理,我们仍然可以把平均增加的百分比当作正态分布处理。
- (b) 替所有公司总裁平均增加的薪水百分比造一个 95% 置信区间。
- (c) 我们必须知道被研究的这 104 公司的何种信息,才能使(b)中所做的推论有根据?

IV.23 水质。某环保组织从一条河随机选取了 45 个定点,在每个定点收集了 1 升河水,并度量了其中的含氧量。所得到的平均数是 4.62 毫克,标准差是 0.92 毫克。这是否为强烈证据,显示整条河的平均含氧量低于每升 5 毫克?

IV.24 宜人的气味。宜人的气味会提高工作效率吗?21 位受试者被要求戴着面具用铅笔在纸上走迷宫,有的面具没有气味,有的会散发花香。每位受试者分别戴两种面具各走 3 次迷宫,面具顺序随机决定(这是配对设计)。以下是他们平均时间(秒)的差,是用没气味减掉有香气的。如果花香会提高工作效率,则这些平均时间的差应该是正



的, 因为有香气的平均时间会较低。

-7.37	-3.14	4.10	-4.40	19.47	10.80	-0.87
8.70	2.94	-17.24	14.30	-24.57	16.17	-7.84
8.60	-10.77	24.97	-4.47	11.90	-6.26	6.67

- (a) 我们想要证明平均来说, 有香气的面具会加快工作进度。用 μ 来表示出原假设及备择假设, μ 是所有成人的总体的平均时间差。
- (b) 用计算机算出 21 个观测值的平均数和标准差。受试者戴香气面具时走迷宫走得比较快吗? 平均缩短的时间长到显示出重要性吗?
- (c) 画一个数据的茎叶图(先四舍五入到秒)。有没有看到异常值或是其他可能妨碍做推论的问题?
- (d) 检验你在(a)中写出的假设, 减少的时间是否具统计显著性?

IV. 25 鲨鱼。大白鲨又大又饿。以下是 44 只大白鲨的长度(单位: 英尺):

18.7	12.3	18.6	16.4	15.7	18.3	14.6	15.8	14.9	17.6	12.1
16.4	16.7	17.8	16.2	12.6	17.8	13.8	12.2	15.2	14.7	12.4
13.2	15.8	14.3	16.6	9.4	18.2	13.2	13.6	15.3	16.1	13.5
19.1	16.2	22.8	16.8	13.6	13.2	15.7	19.7	18.7	13.2	16.8

- (a) 画一个茎叶图, 用英尺当茎, 十分之一英尺当叶。一共有两个异常值, 两个方向各有一个。这样不会影响 \bar{x} , 但是会把标准差 s 拉大。
- (b) 找出所有大白鲨平均长度的一个 90% 置信区间。(因为异常值对 s 的影响, 区间可能会很宽。)
- (c) 我们对这些鲨鱼必须有什么样的了解, 才能够说明(b)中的结果?

IV. 26 宜人的气味。回到习题 IV. 24 的数据。替戴着花香面具走迷宫时省下的平均时间造一个 95% 置信区间。你是不是有信心, 花香有助于减少平均工作时间?

IV. 27 鲨鱼。回到习题 IV. 25 的数据。是不是有合理证据显示, 这



些鲨鱼所代表的总体的平均身长超过 15 英尺?

IV.28 辛浦森悖论。如果我们比较美国全国教育成果评估的平均数学分数, 会发现内布拉斯加州的八年级生考得比新泽西州的八年级生好。但是如果我们只看白人学生的成绩, 则是新泽西州的考得较好。如果我们只看少数族裔的成绩, 新泽西州还是考得比较好。这就是辛浦森悖论: 当我们把两组学生合并考虑时, 结论会倒过来。利用内布拉斯加州八年级生中白人比例高得多的这个事实, 详细说明为什么这种结果不奇怪。

IV.29 不满意的 HMO 病人。一项对参加卫生维护组织 HMO 者所提申诉的研究, 把提出有关医疗申诉的人、提出非关医疗申诉的人以及该年未提申诉者的一个 SRS 做了比较。以下是每组总人数以及自愿退出 HMO 人数的资料。

	未申诉	医疗申诉	非医疗申诉
总人数	743	199	440
退出人数	22	26	28

- 算出每组有多少百分比退出。
- 造一个申诉状况对应是否退出的双向表。
- 求出预期计数并确认符合使用卡方检验的条件。
- 双向表的卡方统计量的值是 $\chi^2 = 31.765$ 。这个统计量检验的原假设和备择假设各是什么? 自由度是多少? 统计显著性如何? 对于申诉状况和退出 HMO 与否之间的关系, 你会做出什么结论?

IV.30 治疗溃疡。胃冷冻曾有一度是胃溃疡的标准疗法, 在实验证明胃冷冻并无疗效之后, 这种疗法就停止使用了。有一项随机化比较实验的结果, 是使用胃冷冻的 82 个病人中有 28 人症状有改善, 而安慰剂组的 78 个病人中, 有 30 人有改善。

- 用图示描绘此实验的设计。
- 造一个疗法对应结果(受试者症状是否改善)的双向表。疗法和结果之间有具统计显著性的相关性吗?
- 写一个简短的总结, 内容要包括检验的结果和用来比较两种疗法成功程度的百分比。



IV.31 什么时候打给你比较好?从习题 IV.11 中我们得知,有一项研究在两个不同时段随机拨打电话。工作日上午所打的 2 304 个电话中,有 1 313 个有人接。工作日晚间所打的 2 454 个电话中,有 1 840 个有人接。

- (a) 造一个时段对应电话是否有人接的双向表。在两个时段中,有人接的百分比各是多少?
- (b) 应该很明显可以看出,时段和电话是否有人接之间,有具高度显著性的相关性,为什么?
- (c) 虽然结果明显,还是把卡方检验做完。结论是什么?



第四部分 报告作业

报告作业是比较长的习题，需要搜集信息或制作数据，而且重点是要把做出的结果用一篇短文来说明，这里很多题目适合由一组学生共同来做。

作业 1. 替医学研究做报告。医学期刊中许多主要文章都和经过统计设计的研究有关，并会报告推论的结果，通常都是 P 值或 95% 置信区间。你可以在《美国医学会期刊》的网站 (jama.ama-assn.org)，或《新英格兰医学期刊》的网站 (www.nejm.org) 上找到当期文章的摘要。要看整篇文章可能要付费，或者得上图书馆。选一篇描述医学实验的文章，实验主题要是没受过医学训练的人也能懂得的，像是 21 章和 22 章中用过的例子：生气和心脏病发作的关系以及食物中的纤维可以降低胆固醇这类。写一篇分成两段的新闻报道，说明该研究的结果。

然后写一篇简短的讨论，说明你怎样决定报道中要写什么和不写什么。比如说，如果你省略了有关统计显著性或者置信区间的细节，要说明理由。关于研究的设计你说了什么，为什么这么说？写新闻报道的人经常要做这类决定。

作业 2. 推论的使用与滥用。有关极复杂统计方法的说明，很少有不需要很多训练就可以读懂的。有一个例子是《伯特的真正错误》，这是古尔德 (Stephen Jay Gould) 在 1981 年由 W. W. Norton 出版的《人的错误量度》 (*The Mismeasure of Man*) 一书中的一章，这个例子也是谈统计推论滥用的极好文章。我们在习题 9.20 中，在可疑的状况下和伯特先生邂逅。古尔德用了很长的一章，指出伯特以及其他专门利用复杂统计去“发现”一些令人起疑的形态。把这一章读一读，并写一段简短说明，解释为什么“因子分析” (factor analysis) 无法对心智能力的结构提供明确的信息。

作业 3. 自己做一项统计研究。找两个类别变量，二者之间的相关关系是你有兴趣的，然后自己去搜集数据。一个简单例子是学生的性别以及他的政治倾向。稍复杂一点的例子是大学生的年级和他毕业后



的计划(马上做事、继续念书、休息一阵子……)。我不坚持你非要用确实的简单随机样本不可。

搜集数据并造一个双向表。然后做分析,包括比较百分比来描述你两个变量间的相关关系,以及用卡方统计量来评估其统计显著性。描述你的研究,以及有什么发现。你的样本会不会太小了,以至于未达到统计显著性都不令人意外?

作业 4. 汽车的颜色?我听说白色的汽车卖得比其他颜色的都多(当然这无疑表示所谓“白色”也包括各种“珍珠白”及“奶油白”)。你学校的学生开的车中有多少百分比是白色?搜集数据并造一个白色汽车所占比例的置信区间来回答这个问题。你可以抽一个学生样本,问他们问题来搜集数据,也可以到学生停车场去看车子的颜色。在你的讨论中,要说明你做了何种努力,来试图得到接近学生汽车的 SRS 的数据。

注释与资料出处

写给教师

第 I—II 页 若想多知道些有关把统计看成文科的想法，可参考我对美国统计学会 (USA) 做的主席演讲：David S. Moore, “Statistics among the liberal arts.” *Journal of the American Statistical Association*, 93 (1998), pp. 1253 – 1259。

第 II 页 此处引用的话摘录自委员会的报告，该报告内容经美国统计学会的理事会全体一致认可。完整的报告是 George Cobb 的 “Teaching statistics”，在 L. A. Steen 所编辑的 *Heeding the Call for Change: Suggestions for Curricular Action*. Mathematical association of America, 1990, pp. 3 – 34。

前言

第 IV 页 前言中所提到的例子，详细内容出现的位置如下：高压电线和儿童白血病：第 1 章例 3。蓝德丝：第 2 章例 2。用胃冷冻治溃疡：第 5 章例 2。赌场及犯罪案件：第 20 章，454 页。教育程度及收入：第 12 章例 3 及 23 章例 4。正常体温：25 章 562 页。度量失业率：第 8 章例 3 及例 10。

第 V 页 汉蒙斯 (Tim Hammonds) 曾说：“社会大众对于对立的意见已习以为常……许多人已经觉得，对应每一位博士，存在一位对等但相反的博士，” 见 *Chance*, 7, No. 2 (1994), p. 9。

第 VII—VIII 页 参考 *Making Statistics More Effective in Schools of Business*, Graduate School of Business, University of Chicago, 1986 中 A. C. Nielsen, Jr. 所著 “Statistics in marketing”。

第 VIII 页 有关乳房 X 光摄影的信息出自 H. C. Sox, “Editorial: benefit and harm associated with screening for breast cancer,” *New England Journal of Medicine*. 338 (1998), p. 1145。

第 IX 页 有关自杀的讨论，参考 Howard I. Kushner, *Self Destruction in the Promised Land*, Rutgers University Press, 1989。



第一部分

第 1 页 Sound Scan 从各商店的记录器搜集音乐销售的电子资料。

许多网站上有各式各样的结果。我在 backstreet.net 看到 1999 年底的数据。

第 1 章

第 2 页 有关投票的数据出自 “Only four in ten Americans ‘always vote’ ”，是由 Lydia Saad 于 1999 年 11 月 2 日提出的盖洛普调查的新闻发布。可上网查，网址：www.gallup.com/poll/releases。

第 4 页 例 1 源自 Maxine Pfannkuch 及 Chris J. Wild 的 “Statistical thinking and statistical practice themes gleaned from professional statisticians,” 1998, 未出版。

第 7 页 例 3 出自 M. S. Linet 等的 “Residential exposure to magnetic fields and acute lymphoblastic leukemia in Children,” *New England Journal of Medicine*, 337(1997), pp. 1–7。

第 9 页 美国社区调查的网站是：www.census.gov/acs/www。

第 9—10 页 对于普查漏掉部分的估计，出自 Howard Hogan 的 “The 1990 post-enumeration survey: operations and results,” *Journal of the American Statistical Association*, 88(1993), pp. 1047–1060。

第 2 章

第 18 页 参考 “Acadian ambulance officials, workers flood call-in poll,” *Baton Rouge Advocate*, 1999 年 1 月 22 日。

第 25 页 例 4：出自 1999 年 11 月 18 日 Mark Gillespie 向媒体公布的盖洛普调查结果：“Majority of smokers want to quit, consider themselves addicted,” 可上网查，网址：www.gallup.com/poll/releases。

第 29 页 习题 2.6 参考 “Pseudo-opinion polls SLOP or useful date?” *Chance*, 8, No. 2(1995), pp. 16–25, D. Horvitz 的部分。

第 3 章

第 34—35 页 参考盖洛普的 “Gambling in America – 1999: a comparison of adults and teenagers” 中 Topline and Trends 的详细报告：可上网阅读，网址：www.gallup.com/poll。



第 46 页 有关对公布选举预测的限制, 信息出自 Richard Morie 的
“Crack down on pollsters,” *Washington Post*, 1998 年 1 月 19 日。

第 51 页 习题 3.7: 参考 Warren MCIIsaac 及 Vivek Goel, “Is access
to physician services in Ontario equitable?” Institute for Clinical Eval-
uative Sciences in Ontario, 1993 年 10 月 18 日。

第 54 页 习题 3.15: 参考 1989 年 8 月 21 日的 *New York Times*。

第 54 页 习题 3.16: 参考 www.louisharris.com 这个网站。它同时也
是习题 3.19 的来源。

第 56 页 习题 3.28: 根据盖洛普网站的一篇报告而来, 网址:
www.gallup.com。

第 4 章

第 58—59 页 参考 Gregory Flemming and Kimberly Parker, “Race and
Reluctant respondents: possible consequences of non-response for
pre-election surveys,” Pew Research Center for the People and the
Press, 1997, 可在 www.peoplepress.org 上找到。

第 62 页 想要多了解些非抽样误差以及相关信息, 可参考
P. E. Converse 及 M. W. Traugott 的 “Assessing the accuracy of polls
and Surveys,” *Science*, 234(1986), pp. 1094 – 1098。

第 62—63 页 想要多了解些抽样调查时受限于受访者记忆造成的影
响, 可参考 N. M. Bradburn, L. J. Rips 及 S. K. Shevell 的 “Answering
autobiographical questions: the impact of memory and inference on Sur-
veys,” *Science*, 236(1987), pp. 157 – 161。

第 64 页 例 4: CPS 的不回应率出自 “Technical notes to household
survey data published in Employment and Earnings,” 可在劳工统计
局的网站找到, 网址: <http://stats.bls.gov/cps/home.htm>。全面社
会调查也在自己的网站上报告了回应率, “已掉到 20% 这么低”
这项声明, 出自 Don Van Natta, Jr. 的 “Polling’s dirty little se-
cret’: no response,” *New York Times*, 1999 年 11 月 11 日。

第 65 页 例 5: 对福利制度的回应出自 1992 年 7 月 5 日, *New York
Times* 所报导的一项 *New York Times*/CBS News 民意调查结果。有
关苏格兰的那段出自 “All set for independence?” *Economist*, 1998
年 9 月 12 日。

第 66 页 例 6: 参考 D. Goleman, “Pollsters enlist psychologists in
quest for unbiased results,” *New York Times*, 1993 年 9 月 7 日。



第 67 页 加权这一段出自 Adam Clymer 及 Janet Elder 的 “Poll finds greater confidence in Democrats,” *New York Times*, 1999 年 11 月 10 日。

第 68 页 例 7: CPS 的设计的最近一份说明是在劳工统计局的 *Design and Methodology*, Current Population Survey Technical Paper 63, 2000 年 3 月。有打印版, 也可上网查看, 网址: www.bls.census.gov/cps/tp/tp63.htm。例 7 的说明省略了许多复杂情况, 比如说有必要对诸如大学学生宿舍等的不同住宿区分别抽样。

第 74 页 习题 4.1: 中间那段话是盖洛普民意调查的典型声明, 可在盖洛普网站: www.gallup.com 找到。

第 74 页 习题 4.4: 参考 1995 年 5 月 29 日, *New York Times*, P. H. Lewis 的 “Technology” 专栏。

第 75 页 习题 4.6: 参考 Giuliana Coccia 的 “An overview of non-response in Italian telephone surveys”, *Proceedings of the 99th Session of the International Statistical Institute*, 1993, Book 3, pp. 271 – 272。

第 75 页 习题 4.7: 参考 Richard Morin 的 “It depends on what your definition of ‘do’ is”, *Washington Post*, 1998 年 12 月 21 日。

第 76—77 页 习题 4.12: 参考 John Simons 的 “For risk takers, system is no longer sacred”, *Wall Street Journal*, 1999 年 3 月 11 日。

第 77 页 习题 4.13: 参考 Adam Clymer 及 Janet Elder 的 “Poll finds greater confidence in Democrats,” *New York Times*, 1993 年 9 月 7 日。

第 80 页 习题 4.23: 出自 G Gaskell 等人的 “Worlds apart? The reception of genetically modified foods in Europe and the U.S.,” *Science*, 285(1999), pp. 384 – 387 的网上补充资料 (Supplementary Material)。

第 5 章

第 84 页 参考 Allan H. Schulman 及 Randi L. Sims 的 “Learning in an online format versus an in-class format: an experimental study,” *T. H. E. Journal*, June 1999, pp. 54 – 56 及 L. L. Miao 的 “Gastric freezing: an example of the evaluation of medical therapy by randomized clinical trials,” 后者收录于 J. P. Bunker, B. A. Barnes 及 F. Mosteller 所编辑的 *Costs, Risks and Benefits of Surgery*, Oxford U-



niversity Press, 1977, pp. 198 - 211. 这也是例 2 的来源。卡罗来纳启蒙计划的细节, 包括已出版的参考资料, 可以在下列网站找到: [WWW.fpg.unc. edu/over view/abc/abc-ov. htm](http://WWW.fpg.unc.edu/over view/abc/abc-ov. htm)。

第 89 页 例 3: 参考 Samuel Charache 等所著, “Effects of hydroxyurea on the frequency of painful crises in sickle cell anemia,” *New England Journal of Medicine*, 332(1995), pp. 1317 - 1322。

第 91 — 92 页 参考 H. Sacks, T. C. Chalmers 和 H. Smith, Jr., “Randomized versus historical controls for clinical trials,” *American Journal of Medicine*, 72(1982), pp. 233 - 240。

第 94—95 页 例 5 参考 Marilyn Ellis 的美联社报道, “Age not bias, may explain differences in treatment,” 刊登于 1994 年 4 月 26 日的 *New York Times*, Dr. Mark 是在评论 Daniel B. Mark 等人的 “Absence of sex bias in the referral of patients for cardiac catheterization,” *New England Journal of Medicine*, 330(1994), pp. 1101 - 1106。要想看到相对的研究评论, 参考 D. Douglas Miller 及 Leslee Shaw 的 “Sex bias in the care of patients with cardiovascular disease,” *New England Journal of Medicine*, 331(1994), p. 883。

第 100 页 习题 5.10: 参考 G. Kolata 的 “New study finds vitamins are not cancer preventers,” *New York Times*, 1994 年 7 月 21 日。若要读细节, 请参考同日出版的 *Journal of the American Medical Association*。

第 100—101 页 习题 5.12: 参考 Stan Metzenberg 的 Letter to the editor, *Science*, 286(1999), p. 2083。

第 101 页 习题 5.14 的根据是 Christopher Anderson 的 “Measuring what works in health care,” *Science*, 263(1994), pp. 1080 - 1082。

第 103 页 习题 5.23: 参考 L. E. Moses 及 F. Mosteller 的 “Safely of anesthetics”。收录于 J. M. Tanur 等人所编辑的 *Statistics: A Guide to the Unknown*, 第 3 版, Wadsworth, 1989, pp. 15 - 24。

第 6 章

第 106—107 页 参考 Martin Enserink 的 “Fickle mice highlight test problems,” *Science*, 284(1999), pp. 1599 - 1600。同一期中亦有该研究的完整报告。

第 108 页 例 1: 关于老鼠的事实出自 E. Street 和 M. B. Carroll 的 “Preliminary evaluation of a new food product”。收录在 J. M. Tanur



- 等人所编辑的 *Statistics: A Guide to the Unknown*, 第 3 版, Wadsworth, 1989, pp. 161 - 169.
- 第 109 页 例 2: 安慰剂效应的例子出自 Sands Blakeslee 的 “Placebos prove so powerful even experts are surprised,” *New York Times*, 1998 年 10 月 13 日。“四分之三”的估计会被 Martin Enserink 引用于 “Can the placebo be the cure?” *Science*, 284(1999), pp. 238 - 240。更详细的讨论在 Anne Hanington 所编辑的 *The Placebo Effect: An Interdisciplinary Exploration*, Harvard University Press, 1997。
- 第 111 页 所引用的流感实验出自 Kristin L. Nichol 等人的 “Effectiveness of live, attenuated intranasal influenza virus vaccine in healthy, working adults,” *Journal of the American Medical Association*, 282 (1999), pp. 137 - 144。
- 第 112 页 例 3: 参考 “Cancer clinical trials: barriers to African American participation,” *Closing the Gap*, newsletter of the Office of Minority Health, December 1997 - January 1998。
- 第 113 页 例 4: 参考 Michael H. Davidson 等人所著 “Weight control and risk factor reduction in obese subjects treated for 2 years with orlistat: a randomized controlled trial,” *Journal of the American Medical Association*, 281(1999), pp. 235 - 242。
- 第 116 页 例 8: 是以下文章的简化版本, Arno J. Rethans, John L. Swasy, 和 Lawrence J. Marks 的 “Effects of television commercial repetition, receiver knowledge, and commercial length: a test of the two-factor model,” *Journal of Marketing Research*, 23(February 1986), pp. 50 - 61。
- 第 118 页 例 9: 参考 “Advertising: the cola war,” *Newsweek*, 1976 年 8 月 30 日, p. 67。
- 第 122 页 习题 6.3: 参考 Edward P. Sloan 等人的 “Diaspirin cross-linked hemoglobin(Dclhb) in the treatment of severe traumatic hemorrhagic shock,” *Journal of the American Medical Association*, 282 (1999), 1857 - 1864。
- 第 124 页 习题 6.10: 参考 W. E. Nelson, R. C. Henderson, L. C. Almekinders, R. A. DeMasi 及 T. N. Taft 的 “An evaluation of pre- and postoperative nonsteroidal antiinflammatory drugs in patients undergoing knee arthroscopy,” *Journal of Sports Medicine*, 21(1994),



pp. 510 - 516.

第 127 页 习题 6.22: 参考 Mary O. Munding 等人的 “Primary care outcomes in patients treated by nurse practitioners or physicians,” *Journal of the American Medical Association*, 238(2000), pp. 59 - 68.

第 7 章

第 128 页 所引用的话出自 Thomas B. Freeman 等人的 “Use of placebo surgery in controlled trials of a cellular-based therapy for Parkinson’s disease,” *New England Journal of Medicine*, 341(1999), pp. 988 - 992. Freeman 支持帕金森氏症实验。反对者的代表是 Ruth Macklin, 参考她的 “The ethical problems with sham surgery in clinical research,” *New England Journal of Medicine*, 341(1999), pp. 992 - 996.

第 130 页 例 1: 参考 John C. Bailar III 的 “The real threats to the integrity of science,” *The Chronicle of Higher Education*, April 21, 1995, pp. B1 - B2.

第 132 页 例 2: 对医学研究的知情且同意及审查委员会工作的规范, 解释起来的困难情况, 是 Beverly Woodward 这篇文章的主题 “Challenges to human subject protections in U. S. medical research,” *Journal of the American Medical Association*, 282(1999), pp. 1947 - 1952. 这篇文章的参考文献中有其他相关讨论。

第 134 页 亨内肯医师的话引用自 Annenberg/Corporation 替公共电视制作的 *Against All Odds: Inside Statistics* 中的一段访问。

第 135 页 例 4: 引用之内容出自 *Report of the Tuskegee Syphilis Study Legacy Committee*, 1996 年 5 月 20 日。更详细的历史要参考 James H. Jones 的 *Bad Blood: The Tuskegee Syphilis Experiment*, Free Press, 1993.

第 137 页 引用内容出自 Gina Kolata 和 Kurt Eichenwald 的 “Business thrives on unproven care leaving science behind,” *New York Times*, 1999 年 10 月 3 日。背景资料及最初做的一些临床试验的细节出现在美国国家癌症研究所的新闻稿 “Questions and answers: high-dose chemotherapy with bone marrow or stem cell transplants for breast cancer,” April 15, 1999. 所报道的研究中有伪造数据的那项, 是由 Denise Grady 报道的 “Breast cancer researcher admits falsifying data,” *New York Times*, 2000 年 2 月 5 日。

第 138 页 例 7: 参考 R. D. Middlemist, E. S. Knowles 及 C. F. Matter 的



“Personal space invasions in the lavatory: suggestive evidence for arousal,” *Journal of Personality and Social Psychology*, 33(1976), pp. 541 – 546.

第 145 页 习题 7.13: 所引用的话是 Mt. Sinai School of Medicine 的 Chief of neurology, Dr. C Warren Olanow 说的, 参考 Margaret Talbot 的 “The placebo prescription,” *New York Times Magazine*, January 8, 2000, pp. 34 – 39, 44, 58 – 60。

第 145 页 习题 7.14: 需要更多背景资料, 可参考 Jon Cohen 的 “AIDS trials ethics questioned,” *Science*, 276(1997), pp. 520 – 523。另外有关习题 7.14, 7.15 及 7.16 的后续争议的最近发展, 可以去网站: www.sciencemag.org 的 archives 部分找。

第 148 页 习题 7.24: 引用部分出自 *Science*, 192(1976), p. 1086。

第 8 章

第 150 页 参考 Janny Scott 的 “Working hard, more or less: studies of leisure time point both up and down,” *New York Times*, 1999 年 7 月 10 日。我是从 *Chance News* 的网站得知这篇文章的, 网址: www.dartmouth.edu/~chance/chance-news

第 151 页 例 1: 参考 “Trial and error,” *Economist*, October 31, 1998, pp. 87 – 88。

第 157 页 例 7: 引用的话出自 1999 年 8 月 31 日 “公正测验” 发布的新闻稿, 可在该组织的网页上找到, 网址为: www.fartest.org。此例及统计争议中大部分其他信息, 出自 National Science Foundation 的报告 *Women, Minorities and Persons with Disabilities in Science and Engineering: 1998*, NSF99 – 338, 1999 的第 2 章。统计学的争议中的表所列出的 r^2 的值, 是根据相关系数算出的, 相关系数可以在 College Board 的网站找到, 网址: www.collegeboard.org。

第 162 页 例 9: NIST 时间和 BIPM 时间的差距, 是从 NIST Time and Frequency Division 的网站下载下来的, 网址为: www.boulder.nist.gov/time_freq。

第 9 章

第 174—175 页 消失的厢型车事件, 出处为 1992 年 4 月 18 日及 9 月 10 日 *New York Times* 的新闻报道。



- 第 176 页 例 1: 参考 R. J. Newan, "Snow job on the slopes," *US News and World Report*, December 17, 1994, pp. 62 - 65.
- 第 177 页 例 3: 参考 "Colleges inflate SATs and graduation rates in popular guide-books," *Wall Street Journal*, 1995 年 4 月 5 日。
- 第 177—178 页 例 4: 参考 Robyn Meredith 的 "Oops, Cadillac says, Lincoln won after all," *New York Times*, 1999 年 5 月 6 日。
- 第 178 页 例 5: 参考 B. Yunker, "The strange case of the painted mice," *Saturday Review / World*, November 30, 1974, p. 53.
- 第 179 页 例 6: 写信的是 J. L. Hoffmam, 参考 *Science*, 255(1992), p. 665 最初的文章出现的日期是 1991 年 12 月 6 日。
- 第 180 页 例 7: 参考 E. Marskall, "San Diego's tough stand on research fraud," *Science*, 234(1986), pp. 534 - 535.
- 第 181 页 例 9: 其中的广告是 Darryl Nester 看到的。
- 第 181—182 页 例 10: 见 *Science*, 192(1976), p. 1081。
- 第 183 页 例 12: 所引用的话出自美国心脏病协会的 "Cardiovascular disease in women"。www.americanheart.org 网站上有。
- 第 184 页 例 13: 参考 Edwin S. Rubenstein 的 "Inequality," *Forbes*, November 1, 1999, pp. 158 - 160 及 Bureau of the Census, *Money Income in the United States*, 1998。
- 第 186 页 习题 9.1: 参考 *Providence Journal* (Rhode Island), December 24, 1999。我是在 *Chance News* 9.92 看到的。
- 第 186 页 习题 9.4: 参考 *Fine Gardening*, September/October 1989, p. 76。
- 第 186—187 页 习题 9.6: 参考 *Conde Nast Traveler* 杂志, June 1992。
- 第 187 页 习题 9.7: 参考 *Science*, 189(1975), p. 373。
- 第 187 页 习题 9.8: 参考 *Lafayette (Ind.) Journal and Courier*, October 23, 1988。
- 第 187 页 习题 9.9: 信是 L. Jarvik 写的, 参考 1993 年 5 月 4 日 *New York Times*。社论 "Muggings in the Kitchen", 出现于 1993 年 4 月 23 日。
- 第 188 页 习题 9.10: 见 *New York Times*, 1986 年 4 月 21 日。
- 第 188 页 习题 9.15: 见 *Purdue Exponent*, 1977 年 9 月 21 日。
- 第 189 页 习题 9.17: 和 9.18 分别参考 1983 年 7 月和 1983 年 3 月的 *Organic Gardening*。



第一部分 复习

第 198 页 习题 I. 10—I. 12, 可在盖洛普网站 www.gallup.com 找到盖洛普调查的资料。习 I. 10 引用的话, 出自 1999 年 8 月 30 日盖洛普发布的新闻稿中, David W. Moore 的 “Americans support teaching creationism as well as evolution in public schools”。

第 199 页 习题 I. 16 到 I. 19: 参考 Kristin L. Nichol 的 “Effectiveness of live, attenuated intranasal influenza virus vaccine in healthy, working adults,” *Journal of the American Medical Association*, 282(1999), pp. 137–144。

第二部分

第 10 章

第 214 页 例 5: 图 10. 7 的价格, 结合了无铅普通汽油的 CPI Component(是一种指数)以及 Energy Information Administration 于 1999 年 1 月所报告的实际价格。这项信息在网上找得到。

第 226 页 习题 10. 16: 参考 1992 年 7 月 15 日 *New York Times*, E. Norris 的 “Market Watch” 专栏。

第 227 页 习题 10. 18: 参考 Bureau of Economic Analysis 的 *Survey of Current Business* 在 www.bea.doc.gov 网站上有。

第 227 页 习题 10. 21: 参考 1997 *Statistical Yearbook of the Immigration and Naturalization Service*, U. S. Department of Justice, 1999。

第 229 页 习题 10. 25: 参考 Centers for Disease Control and Prevention 的 *National Vital Statistics Reports*, June 30, 1999。

第 11 章

第 233 页 图 11. 1 及 11. 2: 1999—2000 的学杂费资料出自 College board 的 *Annual Survey of Colleges*, 1999–2000。

第 241 页 图 11. 5: 参考 C. B. Williams 的 *Style and Vocabulary: Numerical Studies*, Griffin, 1970。

第 245 页 图 11. 7 是根据 Annenberg/Corporation 为公共电视制作的课程 *Against All Odds: Inside Statistics* 中的一集。



第 246 页 图 11.9: 数据来自政府的 Survey of Earned Doctorates, 可在网上 supplement 找到, 对 Jeffery Mervis 的 “Wanted: a better way to boost numbers of minority PhDs,” *Science*, 281(1998), pp. 1268 – 1270.

第 247 页 图 11.10: 根据 J. K. Ford, “Diversification: how many stocks will suffice?” *American Association of Individual Investors Journal*, January 1990, pp. 14 – 16.

第 247—248 页 表 11.2: 参考 Environmental Protection Agency 的 *Model Year 2000 Fuel Economy Guide*, 1999.

第 248 页 表 11.3: 参考 Ali H. Mokdad 等人的 “The spread of the obesity epidemic in the United States, 1991 – 1998,” *Journal of the American Medical Association*, 282(1999), pp. 1519 – 1522.

第 250 页 习题 11.13: 参考 *Consumer Reports*, June 1986, pp. 366 – 367。稍后的一项热狗研究结果见 1993 年 7 月的 *Consumer Reports*, pp. 415 – 419。较新的数据还涵盖了少数品牌的家禽肉热狗, 而且卡路里计数主要是从包装上读到的, 结果出现令人起疑的一些整数。

第 252 页 表 11.6: 数据出自 National Oceanic and Atmospheric Administration 的网站: www.ncdc.noaa.gov。

第 12 章

第 254—255 页 关于收入和教育程度的数据出自 1999 年 3 月的 CPS Supplement, 是用政府表的 FERRET 系统下载自 Census Bureau 网站。这些是原始调查结果, 包括许多没有收入, 甚至收入为负的人。为了方便比较, 我只报告了四种教育程度。

第 262—263 页 图 12.4 和 12.5 的数据, 出自 U. S. Bureau of the Census 的 *Money Income in the United States, 1998, 1999*。表 B-2 及 B-3 的详细讨论可参考 John H. Hinderaker Scott W. Johnson 的 “The truth about income inequality,” Center of the American Experiment, 1995 (www.amexp.org/publications/)。他们引用了对所得税申报的研究。对儿童的研究参考 Peter Gottschalk 及 Sheldon Danziger 的 “Income mobility and exits from poverty of American children, 1970 – 1992,” Boston University Department of Economics Working Paper 430, 1999。图 12.6 的数据出自 U. S. Bureau of the Census, *Poverty in the United States, 1998, 1999*。



第 268 页 表 12.1 的薪水是从数个网站得来的估计。大部分球员都有复杂的多年合同。

第 273 及 277 页 习题 12.9 及 12.21: 数据来源为 *Consumer Reports*, June 1986, pp. 366 – 367。

第 274 页 习题 12.13: 数据来源为 1999 *Statistical Abstract of the United States*。

第 275 及 276 页 习题 12.15 及 12.19 后的原始数据, 来源是 College Board Online, www.collegeboard.org。

第 276—277 页 习题 12.20: 参考 Environmental Protection Agency, *Model Year 2000 Fuel Economy Guide, 1999*。

第 279 页 习题 12.27: 此题数据是表 15.12 中数据的一部分, 参考习题 15.17 的来源。

第 279 页 习题 12.30: 引用的话出自 *New York Times*, 1989 年 5 月 31 日。

第 13 章

第 300 页 图 13.13 的 IQ 分数, 是由普度大学教育学院的 Darlene Gordon 所搜集的。

第 301 页 习题 13.10: 参考 Stephen Jay Gould, “Entropic homogeneity isn’t why no one hits .400 anymore,” *Discover*, August 1986, pp. 60 – 66。

第 303—304 页 习题 13.22: 参考 Ulric Neisser 的 “Rising scores on intelligence test,” *American Scientist*, September-October 1997 网络版。

第 14 章

第 307 及 308 页 图 14.1 和 14.2 的 1998 年资料, 出自 College Board 网页: www.collegeboard.org。

第 310 页 例 1: 数据出自 World Bank 的 1999 *World Development Indicators* 预期寿命是 1977 年的估计, 每人平均国内生产毛值(以相同购买力为基础)是 1998 年的资料。

第 312 页 例 2: 参考 “Allometric scaling in the earliest fossil bird, *Archaeopteryx lithographica*,” *Science*, 247(1990), pp. 195 – 198。作者根据多项证据做了结论, 所有样本代表同一种类。

第 325—326 页 习题 14.22: 引用的话出自 *T. Rowe Price Report*,



1997 年冬, P4。

第 326 页 习题 14.23: 参考 *Philadelphia City Paper*, May 23 – 29, 1997。因为汽水有大有小, 我把汽水价格依每盎司价格转换成 16 盎司的汽水价格。

第 327 页 习题 14.25: 参考 W. L. Colville 及 D. P. McGill 的 “Effect of rate and method of planting on several plant characters and yield of irrigated corn, ” *Agronomy Journal*, 54(1962), pp. 235 – 238。

第 15 章

第 331 页 引用的话出自 Galatea Corporation 网站: www.voicenet.com/~mitochon。

第 342 页 统计学上的争议中提到的一些研究包括 John R. Lott Jr. 的 *More Guns, Less Crime: Understanding Crime and Gun Control Laws*, University of Chicago Press, 1998; Andres Villaveces 等人的 “Effect of a ban on carrying firearms on homicide rates in 2 Colombian cities,” *Journal of the American Medical Association*, 283(2000), pp. 1205 – 1209; 也可参考同一期中 Lawrence W. Sherman 写的社论。还有 Lawrence W. Sherman, James W. Shaw 及 Dennis P. Rogan 的 “The Kansas City gun experiment,” National Institute of Justice, 1995。

第 343 页 例 7: 参考 Laura L. Calderon 等人的 “Risk factors for obesity in Mexican-American girls: dietary factors, anthropometric factors, physical activity, and hours of television viewing, ” *Journal of the American Dietetic Association*, 96(1996), pp. 1177 – 1179。

第 349—350 页 习题 15.7: 数据出自 University of California, San Diego 的 M. H. Criqui, 登在 1994 年 12 月 28 日的 *New York Times*。

第 350 页 习题 15.8: 数据是从一个图估计出来的, 图在 G. D. Martinsen, E. M. Driebe 和 T. G. Whitham 的 “Indirect interactions mediated by changing plant chemistry: beaver browsing benefits beetles, ” *Ecology*, 79(1998), pp. 192 – 200。

第 351—352 页 习题 15.13: 参考 W. M. Lewis 和 M. C. Grant 的 “Acid precipitation in the western United States,” *Science*, 207(1980), pp. 176 – 177。

第 355 页 习题 15.25: 引用的话出自 Gannett News Service 的一篇文章, 登载在 *Lafayette(Ind.) Journal and Courier*, 1994 年 4 月 23 日。



第 356 页 习题 15.29 的交响乐团广告, 是由 Kalamazoo College 的 Marigene Arnold 提供的。

第 16 章

第 358—359 页 引述的话出自 Bureau of Labor Statistics, “Updated response to the recommendations of the advisory commission to study the Consumer Price Index,” June 1998. BLS *Monthly Labor Review* 1996 年 12 月那期的整本内容都在谈 1998 年 1 月生效的新版 CPI 的主要不同之处。下面这篇文章讨论了 CPI 改变之后的影响: Kenneth J. Stewart 和 Stephen B. Reed 的 “CPI research series using current methods, 1978 - 98,” *Monthly Labor Review*, 122(1999), pp. 29 - 38。以上这些全部可以在 BLS 的网站上找到。

第 370 页 有关政府统计调查的报告下面文章中有: “The good statistics guide,” *Economist*, September 13, 1993, p. 65。

第 380 页 习题 16.25: 参考 Gordon M. Fisher 的 “Is there such a thing as an absolute poverty line over time? Evidence from the United States, Britain, Canada, and Australia on the income elasticity of the poverty line,” U. S. Census Bureau Poverty Measurement Working Papers, 1995。

第 380 页 习题 16.26: 参考 G. J. Borjas 的 “The internationalization of the U. S. Labor market and the wage structure,” *Federal Reserve Bank of New York Economic Policy Review*, 1, No. 1(1995), pp. 3 - 8, 引用的内容出现在第 3 页。习题里面的文章都在试图解释收入为何停滞不前以及收入的差距。大家一致的意见是: 我们不知道。

第二部分 复习

第 390 页 图 II.3 所画的数据出自 G. A. Sacher 和 E. F. Staffelt 的 “Relation of gestation time to brain weight for placental mammals: implications for the theory of vertebrate growth,” *American Naturalist*, 108 (1974), pp. 593 - 613。我是在下面这本书里找到的: Fred L. Ramsey and Daniel W. Schafer 的 *The Statistical Sleuth: A Course in Methods of Data Analysis*, Duxbury, 1997, p. 228。

第 393 页 习题 II.24: 参考 Antoni Basinski 的 “Almost never on Sunday: implications of the patterns of admission and discharge for common conditions,” Institute for Clinical Evaluative Sciences in Ontario,



October 18, 1993。

第 394—395 页 习题 II. 28 最后的正式选举结果得自 Federal Election Commission 的网站: www.fec.gov。

第 395 页 习题 II. 29: 到 1999 年 12 月为止的 36 个月, 每月获利的相关系数, 在 2000 年 1 月的 *Fidelity Insight* 的 newsletter 中有。

第三部分

第 17 章

第 400—401 页 费曼关于航天飞机失败概率的发言, 可在以下他报告的附录中找到: Presidential Commission report “Personal Observations on the Reliability of the Shuttle by R. P. Feynman”、参考 Kennedy Space Center 的网站: www.ksc.nasa.gov/shuttle/missions/51-1/docs/rogers-commission/Appendix-F.txt。

第 404—405 页 更多历史背景可以在下面这本书的头几章找到: F. N. David 的 *Games, Gods and Gambling*, Charles Griffin and Co., 1962。此处提到的历史, 就是出自这本很精彩又有趣的书。

第 405—406 页 想多读些讨论及有趣的例子, 可参考 A. E. Watkins 的 “The law of averages, ” *Chance*, 8, No. 2(1995), pp. 28 – 22。

第 406—407 页 例 5 参考 R. Vallone 和 A. Tversky 的 “The hot hand in basketball: on the misperception of random sequence, ” *Cognitive Psychology*, 17(1985), pp. 295 – 314。

第 408 页 例 7 沃本镇的相关信息参考 S. W. Lagakos 和 B. J. Wessen 以及 M. Zelen 的 “An analysis of contaminated well water and healthy effects in Woburn, Massachusetts, ” *Journal of the American Statistical Association*, 81(1986), pp. 583 – 596, 以及后面接着的讨论。有关兰道夫的信息, 参考 R. Day, J. H. Ware, D. Wartenberg 和 M. Zelen 的 An investigation of a reported cancer cluster in Randolph, Ma. , ” Harvard School of Public Health Technical Report, June 27, 1988。

第 410—412 页 把个人概率和长期比例当做不同概念来呈现的这种做法, 是受了心理学科研究的影响, 这类研究似乎显示出, 人们对于单独问题的判断, 是和分析相对次数这类问题不一样的。至少在 Tversky 和 Kahneman 的经典著作中所出现, 有关我们对机遇的看的部分偏差, 在向受试者说清楚用哪种概率解释后似乎就



消失了。这是很复杂的领域而本书是简单的书，不过我觉得太强调 Tversky 和 Kahneman 的看法的话，可能有点过时了。参考 Gerd Gigerenzer 的 “How to make cognitive illusions disappear: beyond heuristics and biases,” in Wolfgang Stroebe and Miles Hewstone (eds) *European Review of Social Psychology*, Volume 2, Wiley, 1991, pp. 83 – 115。

第 412 页 例 9: 估计的概率出自 R. D’ Agostino, Jr. 和 R. Wilson 的 “Asbestos: the hazard, the risk, and public policy,” 收在 K. R. Foster, D. E. Bernstein 及 P. W. Huber 所编辑的 *Phantom Risk: Scientific Inference and the Law*, MIT Press, 1994, 183 – 210。也可参考 B. T. Mossman 等人所做的类似结论 “Asbestos: scientific developments and implications for public policy,” *Science*, 247(1990), pp. 294 – 301。

第 412 页 引用的话出自 R. J. Zeckhauser 和 W. K. Viscusi 的 “Risk within reason,” *Science*, 248(1990), pp. 559 – 564。

第 417 页 习题 17.13: 参考 T. Hill 的 “Random-number guessing and the first digit phenomenon,” *Psychological Reports*, 62(1988), pp. 967 – 971。

第 19 章

第 436—437 页 起头处的个案研究，部分根据 Nico M. van Dijk 等的 “Designing the Westerscheldetunnel toll plaza using a combination of queuing and simulation,” 收在 P. A. Farrington 等人所编辑的 *Proceedings of the 1999 Winter Simulation Conference*, INFORMS, 1999, pp. 1272 – 1279。

第 450 页 习题 19.13 习题 19.14: 随机甲虫或许在昆虫学界不有名，在模拟世界里却是很有名的。有人说这种甲虫是由 School Mathematics Study Group 的 Arthur Engle 发明的。

第 20 章

第 454—455 页 Andrew Pollack 的 “In the gaming industry, the house can have bad luck, too,” *New York Times*, July 25, 1999。这篇文章中提到 Mirage Resorts 这家赌场一部分因为在百家乐纸牌游戏中的运气不佳，而提出当季利润有问题的警讯。

第 462 页 有关统计学上的争议的内容，可参考“网络寻奇”所提到



的网站，以及“Gambling on the future,” *Economist*, June 26, 1999。有关穷人的论点是根据 NGISC 工作人员的乐透彩报告，参考 www.ngisc.gov/research。

第 466 页 习题 20.6: 访问发布者是 Karen Freeman, *New York Times*, 1996 年 6 月 6 日。

第 466 页 习题 20.7: 根据 A. Tversky and D. Kahneman 的“Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment,” *Psychological Review*, 90(1983), pp. 293 – 315。

第三部分 复习

第 479 页 习题 III. 13 出自 2000 年的一项民意调查，由 National Center for Public Policy and Higher Education 主办，可查阅网站：www.highereducation.org。

第四部分

第 21 章

第 484—485 页 参考 Janice E. Williams 等人的“Anger proneness predicts coronary heart disease risk,” *Circulation*, 101(2000), pp. 2034 – 2039。

第 486 页 例 1 参考 Joseph H. Catania 的“Prevalence of AIDS-related risk factors and condom use in the United States,” *Science*, 258 (1992), pp. 1101 – 1106。

第 489 页 要知道 p 的置信区间的最新发展，可参考 Alan Agresti 和 Brent Coull 的“Approximate is better than ‘exact’ for interval estimation of binomial proportions,” *American Statistician*, 52(1998), pp. 119 – 126。两位作者指出，我们替 p 选的置信区间的精确程度，只要“加上 2 个成功及 2 个失败”就可大大改善。也就是说，把 \hat{p} 用 $(\text{成功个数} + 2) / (n + 4)$ 来取代。抽样调查的教科书里，有把总体为有限这个事实列入考虑之后得出的置信区间，也有在比 SRS 复杂的抽样设计之下的置信区间。

第 500 页 习题 21.12: 参考 Laurie Goodstein 和 Marjorie Connelly 的“Teenage poll finds support for tradition,” *New York Times*, April 30, 1998。



第 502 页 习题 12.21: 参考 Sara J. Solnik 和 David Hemenway 的
“Complaints and disenrollment at a healthy maintenance organiza-
tion,” *Journal of Consumer Affairs*, 26(1992), pp. 90 – 103。

第 22 章

第 504—505 页 这是 Reynolds Farley 等人在下面这篇文章中所谈到的研究的一部分, 但稍微经过精简: “Stereotypes and segregation: neighborhoods in the Detroit area,” *American Journal of Sociology*, 100(1994), pp. 750 – 780。图 22.1 是根据本篇文章第 754 页一个图的一部分得来的。

第 516 页 习题 22.3: 参考 Manisha Chandalia 等人的 “Beneficial effects of high dietary fiber intake in patients with type 2 diabetes mellitus,” *New England Journal of Medicine*, 342(2000), pp. 1392 – 1398。

第 516 页 习题 22.4: 参考 Arthur Schatzkin 等人的 “Lack of effect of a low-fat, high-fiber diet on the recurrence of colorectal adenomas,” *New England Journal of Medicine*, 342(2000), pp. 1149 – 1155。

第 516—517 页 习题 22.5 Fekri A. Hassan 的 “Burials, pig, and political prestige in neolithic China,” *Current Anthropology*, 35(1994), pp. 119 – 141。

第 517 页 习题 22.6: 参考 Fekri A. Hassan 的 “Radiocarbon Chronology of predynastic Nageda settlements, upper Egypt,” *Current Anthropology*, 25(1984), pp. 681 – 683。

第 517 页 习题 22.7: 参考 Sara J. Solnick and David Hemenway 的 “The deadweight loss of Christmas: comment,” *American Economic Review*, 86(1996), pp. 1299 – 1305。

第 521 页 习题 22.21: 参考 Laurie Goodstein 和 Marjorie Connelly 的
“Teenage poll finds support for tradition,” *New York Times*, April 30, 1998。

第 521 页 习题 22.23 出自 Eric Dsslander 的 letter to the editor, *Science*, 257(1992)p. 1461。

第 521 页 习题 22.24 N. Teed, K. L. Adrian 和 R. Knoblauch 的 “The duration of speed reductions attributable to radar detectors,” *Accident Analysis and Prevention*, 25(1991), pp. 131 – 137。这是 Electionic Encyclopedia of Statistical Examples and Exercises (EESEE) 案例研究之一。



第 23 章

第 522—523 页 2000 年 5 月上旬, 我访问 *Smart Money* 网站 www.smartmoney.com 檢視了排名最前的一些基金。最近一篇典型研究, 即 Mark Carhart 的 “On persistence in mutual fund performance,” *Journal of Finance*, 52(1999), pp. 57–82 的部分摘要如下: “根据一个不会有现存偏差的样本, 我证明了股票获利及投资成本的共同因素, 几乎可以完全解释股票共同基金在经过风险调整的平均获利, 为何可保持不变……惟一尚不能解释的重要持续性, 主要是在获利最差共同基金的糟糕表现。”这些结果等于说不需要有信息丰富、技巧高超的共同基金操盘者。

第 530 页 罗森塔尔说的话出现在 B. Azar 的 “APA statistics task force prepares to release recommendations for public comment,” *APA Monitor Online*, 30(May 1999)。网址: www.apa.org/monitor。专门调查委员会的报告参考 Leland Wilkinson 等人的 “Statistical methods in psychology journals: guidelines and explanations,” *American Psychologist*, 54(August 1999), 它提供了合理的应用统计所包含的因素的摘要。

第 535 页 习题 23.4: 参考 Ross M. Stolzenberg 的 “Educational continuation by college graduates,” *American Journal of Sociology*, 99(1994), pp. 1042–1077。

第 24 章

第 543 页 例 2: 和这个例子精神类似的实际数据的例子, 下面文章里面, P. J. Bickel 和 J. W. O'Connell 的 “Is there a sex bias in graduate admissions?” *Science*, 187(1975), pp. 398–404。

第 545 页 例 3: Nationsbank 的资料出自 S. A. Holmes 的 “All a matter of perspective,” *New York Times*, 1995 年 10 月 11 日。

第 546 页 例 4: 参考 D. M. Barnes 的 “Breaking the cycle of addiction,” *Science*, 241(1988), pp. 1029–1030。

第 548—549 页 有很多人用电脑来研究 χ^2 卡方临界值的精确程度。在 Ralph B. D'Agostino 及 Michael A. Stephens 所编辑的 *Goodness-of-Fit Techniques*, Marcel Dekker, 1986, pp. 63–95 的 3、2、5 节 David S. Moore 的 “Tests of chi-square type” 中有关的简短讨论和相关文献。

第 552—553 页 例 7: 参考 Janice E. Williams 等的 “Anger proneness predicts coronary heart disease risk,” *Circulation*, 101(2000),



pp. 2034 – 2039。

第 556 页 习题 24.1 参考 Richard M. Felder 等的 “Who gets it and who doesn't: a study of student performance in an introductory chemical engineering course, ” *1992 ASEE Annual Conference Proceedings*, American Society for Engineering Education, 1992, pp. 157 – 180。

第 556 页 习题 24.2 参考 S. V. Zagona 所编辑的 *Studies and Issues in Smoking Behavior*, University of Arizona Press, 1967, pp. 157 – 180。

第 556 页—557 页 习题 24.3: 参考 R. Shine, T. R. L. Madsen 和 M. J. Elphick 以及 P. S. Harlow 的 “The influence of nest temperatures and maternal brooding on hatchling phenotypes in water pythons, ” *Ecology*, 78(1997), pp. 1713 – 1721。

第 557 页 习题 24.4: 参考 S. W. Hargarten 等的 “Characteristics of firearms involved in fatalities, ” *Journal of the American Medical Association*, 275(1996), pp. 42 – 45。

第 557 页 习题 24.5: 参考 *Digest of Education Statistics 1997*, 可在 National Center for Education Statistics 网站找到: <http://www.ed.gov/NCES>。

第 558 页 习题 24.6 参考 Francine D. Blau 和 Marianne A. Feber 的 “Career plans and expectations of young women and men, ” *Journal of Human Resources*, 26(1991), pp. 581 – 607。

第 558—559 页 习题 24.8: 出自航空公司给交通部的报告, 在下面文章中有: A. Barnett 的 “Hom numbers can trick you, ” *Technology Review*, October 1994, pp. 38 – 45。

第 559 页 习题 24.9: 参考 M. Radelet 的 “Racial characteristics and imposition of the death penalty, ” *American Sociological Review*, 46(1981), pp. 918 – 927。

第 560 页 习题 24.14: 参考 Brenda C. Coleman 的 “Study: heart attack risk cut 74% by stress management”, 美联社稿, 发表在 *Lafayette (Ind.) Journal and Courier*, October 20, 1997。

第 561 页 习题 24.15: 参考 David M. Blau 的 “The child care labor market, ” *Journal of Human Resources*, 27(1992), pp. 9 – 39。

第 25 章

第 562—563 页 体温研究是 PA. Mackowiak, S. S. Wasserman 和 M. M. Levine 的 “A critical appraisal of 98.6 degrees F, the upper



limit of normal body temperature, and other legacies of Carl Reinhold August Wunderlich, " *Journal of the American Medical Association*, 268(1992), pp. 1578 - 1580。我用的文献, 以及根据原始文章的图所得到的数据, 则是出自 Allen L. Shoemaker 的 "What's normal? Temperature, gender, and heart rate." *Journal of Statistics*, 1996。(这份电子期刊的网址为: www.amstat.org/publications/jse/。)

第 567 页 例 2: 关于 NAEP 测验信息出自 Francisco L. Rivera-Batiz 的 "Quantitative literacy and the likelihood of employment among young adults," *Journal of Human Resources*, 27(1992), pp. 313 - 328。

第 573 页 习题 25.3: 数据由普度大学的 Darlene Gordon 提供。

第四部分 复习

第 582 页 习题 IV.1、IV.2、IV.5 用到盖洛普调查的结果, 可在网站: www.gallup.com 找到。

第 583—585 页 习题 IV.9 及 IV.14: 参考 Matthew K. Wynia 等的 "Physician manipulation of reimbursement rules for patients," *Journal of the American Medical Association*, 283(2000), pp. 1858 - 1865。

第 583—584 页 习题 IV.10: 参考 Jane E. Brody 的 "Alternative medicine makes inroads," *New York Times*, April 28, 1998。

第 584 页 习题 IV.13: 参考 Michael F. Weeks, Richard A. Kulka 和 Stephanie A. Pierson 的 "Optimal call scheduling for a telephone survey," *Public Opinion Quarterly*, 51(1987), pp. 540 - 549。

第 585 页 习题 IV.15: 参考 C. Kirk Hadaway, Penny Long Marler 和 Mark Chaves 的 "What the polls don't show: a closer look at U. S. church attendance," *American Sociological Review*, 58(1993), pp. 741 - 752。

第 586 页 习题 IV.22: 参考 Charles W. L. Hill and Phillip Phan 的 "CEO tenure as a determinant of CEO pay," *Academy of Management Journal*, 34(1991), pp. 707 - 717。

第 586—587 页 习题 IV.24: 参考 A. R. Hirsch 和 L. H. Johnston 的 "Odors and learning," *Journal of Neurological and Orthopedic Medicine and Surgery*, 17(1996), pp. 119 - 126。我的数据是在 Electronic Encyclopedia of Statistical Examples and Exercises(ESEEE), W. H. Freeman, 2000 的个案研究找到的。

第 587 页 习题 IV.25 的鲨鱼数据是 Chris Olsen 提供的, 这是他在潜水



杂志上看到的。

第 588 页 习题 IV.28 的例子，是我在 1999 年 *Chance*, 12, No. 2, pp. 43–44 的 Howard Wainer 专栏“Visual revelations”中找到的。

第 588 页 习题 IV.29：参考 Sara J. Solnick 和 David Hemenway 的“Complaints and disenrollment at a health maintenance organization,” *Journal of Consumer Affairs*, 26(1992), pp. 90–103。

第 588—589 页 习题 IV.30：参考 L. L. Miao 的“Gastric freezing: an example of the evaluation of medical therapy by randomized clinical trials”，收在 J. R. Bunker, B. A. Barnes, and F. Mosteller 所编辑的 *Costs, Risks and Benefits of Surgery*, Oxford University Press, 1977, pp. 198–211。

部分习题解答

同学请注意：这些解答只能当作指导，或者让你对答案用，并不能当作完整的解答。虽然解答中没有列出说明或者计算细节，但你做习题时却一定要包含这些，才算完整答案。

第 1 章

1.1 (a) 车子。(b) 厂牌、车型、车种、排档种类、汽缸数、城市耗油及高速路耗油。最后三项是数值的。

1.3 比如说，住户是否使用资源回收桶。

1.5 对总统工作上的表现是否满意；成年美国居民；被访问的 1210 位成人。

1.7 (a) 美国的成年居民。(b) 整批货的所有木材。(c) 所有美国住户。

1.9 抽样调查

1.11 (a) 是由医师(而非研究者)替每位病人选择治疗方式。(b) 应该用随机方式决定每个病人应接受的处理。

1.13 (a) 不是；受试者是依据健康状况分组的。(b) 总体有可能是企业主管、领导者或成人。变量：健康程度及领导能力。

1.15 (a) 抽样调查。(b) 实验。(c) 观测研究。

第 2 章

2.1 自发性回应样本对总体的代表性可能不够。

2.3 72% 太高了。

2.5 可能的答案：(a) 打电话回应样本。(b) 在学生活动中心门口访问正要进去的学生。

2.7 编号 01—28，选出 04、10、17、19、12、13。

2.9 (a) 编号从 001 到 440，(b) 381、262、183、322、341、185、414、273、190、325、330、029、079、078、118、209、354、239、421、426、435、437、193、099、224。

2.11 (a) 编号从 0001 到 3478。(b) 2940、0769、1481、2975、1315。

2.13 (a) 错。(b) 对。(c) 错。



2.15 接电话的人和不接电话的人或许有不同的特征, 样本中两类人都应该有。

2.17 (a) 随机选择看来合理。(b) 这里并不用随机决定方式。(c) 随机方式应是最好的选择。

第3章

3.1 统计量。

3.3 统计量, 参数。

3.5 (a) 选 19、22、39、50、34; $\hat{p}=0.6$ 。(b) 0.4、0.6、0.4、0、0.6、0.2、0.8、0.8、0.6。(d) 4 个样本的 $\hat{p}=0.6$ 。

3.7 (a) 总体: 安大略省居民; 样本: 受访的 61 239 位。(b) 这么大的样本, 对于男性和女性都应该颇具代表性。

3.9 57% 不会叫人惊讶, 但 37% 就会。

3.11 (a) 高偏差、高变异性。(b) 低偏差、低变异性。(c) 低偏差、高变异性。(d) 高偏差、低变异性。

3.13 样本大小为 100 时的误差界限, 是 25 时的一半。

3.15 (a) 44%—50%。(b) 样本结果可能和总体的值不同。(c) 所用的方法有 95% 的时候会得出正确结果。

3.17 样本大小比较小, 误差界限就会较大。

3.19 (a) 约 706(或在 702—711 之间)。(b) 得到该结果所使用的方法, 通常是会落在真正值的正负 3 个百分点之内。

3.21 0.031。

3.23 (a) 0.030(约 3%)。(b) 我们有 95% 信心, 57%—63% 之间的成人相信有地狱。

3.25 样本大小要变成 4 倍, 到 $n=4\ 036$ 。

3.27 比较大。

3.29 直方图的中心看来是在 20%; 样本中未工作的学生人数从 1(4%) 到 9(36%)。在 50 个样本中, 有 32 个(64%) 有 4、5 或 6 个未工作的学生。

第4章

4.1 比如说: 涵盖不全或者无回应。

4.3 (a) 抽样误差。(b) 非抽样误差。(c) 抽样误差。

4.5 此为(随机)抽样误差, 误差界限已把它考虑进去了。



4.7 “对被指控的事项抗争”和“继续总统职务”可能引起不同的情绪反应。

4.9 第二种措辞中列出的各项计划，会让回应者比较不倾向减税。

4.13 调查由《纽约时报》及CBS新闻执行。样本是从成年美国居民中随机抽出，包括1162个人，从11月4日到7日之间(1999年)用电话联系。但是没报告回应率，也没说出问题的确实问法。

4.15 (a) 最后一组应该最准，因为允许不具名会增加诚实程度。

4.17 每位学生有1/10机会受访，但每个样本必包括恰好3位21岁以上的学生及2位21岁以下的。

4.19 (a) 给男性编号：0001，…，2000；女性：001，…，500，然后使用随机数字表两次。头5位女性是138、159、052、087、359。头5位男性是1369、0815、0727、1025、1868。(b) 男性0.1；女性0.4。

4.21 误差界限随样本大小而变，而不是总体大小。

4.23 (a) 抽样本的过程分成好几个阶段，可能类似当“前人口调查”的样本。(b) “层”是指17个国家。(c) 选样本过程中至少有一个阶段用到随机选择。

4.25 选第35、75、115、155以及第195间。

4.27 (a) 编码0001—3500，然后用随机数字。(b) 先在头14个学生中随机选一位，然后选该学生之后第14、28、42……个。(c) 从两组分别选大小为200及50的SRS。

4.29 在购物中心进行的访谈，其对象并非随机选自任何(有意义的)总体；他们只能代表会去购物中心购物的人。

第5章

5.1 (a) 解释变量：治疗方法；反应变量：存活的时间。(b) 女性自己(或其医师)选择治疗方法。(c) 医师有可能在某种程度上是视病的严重程度来建议疗法。

5.3 母亲之前对婴儿的态度是潜在变量，这点也可能影响对哺乳方式的选择；之前的态度会和哺乳的效果有交叉。

5.5 (a) 受试者：医师；解释变量：药物种类(阿司匹林或安慰剂)；反应变量：健康情况(有没有发作心脏病)。

5.7 我们永远没法知道，态度改变有多少是因为解释变量(读宣传资料)，又有多少是因为那时候发生的历史事件。

5.9 应把修课的学生随机指派到教室学习或上网学习。然后就可以



用一项标准测验的分数来做比较。

5.11 如果今年和去年有些不同(比如说,天气比较暖和),我们就无法比较耗电量。这种差异会和处理的效果混在一起。

5.13 (a) 是一对一的封装衬条。(b) 解释变量:执钳温度(250°F、275°F、300°F、325°F)。(c) 反应变量:拉开封口要用的力。

5.15 一次选两个数字;第一组是 16、04、19、07、10;第二组是 13、15、05、09、08;第三组是 18、03、01、06、11。其他是第四组。

5.17 血压的差异大到不像是因为机遇产生的(如果钙无效的话)。

5.19 (a) 替所有受试者量血压,然后随机选择一半人给他们补充钙,另一半给安慰剂。过一段时间后观察血压的变化。(b) 编号 01—40,选出 18、20、26、35、39、16、04、21、19、37、29、07、34、22、10、25、13、38、15、05。

5.21 除了不具体的治疗方式以外,其他因素的影响已被消除或解释,所以在受试者之间观察到的改进状况,可以归因于处理的差异。

5.23 病人疾病的本质及严重程度,以及病人的整体健康状况,可能会左右麻醉剂的选择,并影响死亡率。我们应考虑手术种类以及病人的年龄、性别及身体状况。

第 6 章

6.1 给予某些受试者安慰剂,可以让研究者观察到参与实验可得到的改进或益处。“双盲”表示不论受试者还是工作人员都不知谁参与何种处理;这样做可以防止研究者因预期心理而影响了对受试者的判断。

6.3 病人不知自己得到什么处理,但治疗他们的人知道。这是维护病人安全的必需的作法。

6.5 比如说,大鼠在短时间内高度暴露以致得癌症,代不代表人类在较长的时间内暴露于较低剂量也会得癌症?

6.7 (a) 安慰剂效应。(b) 用一项包含 3 种处理的完全随机化设计。(c) 不该。(d) 因为是病人自己评估处理的效果,所以不需要双盲设计。

6.9 安慰剂真的可以缓解疼痛。

6.11 (a) 个体:小公鸡;反应变量:增加的体重。(b) 两个解释变数(玉米品种及蛋白质含量);9 种处理。图必须是 3×3 的表。需要 90 只小鸡。(c) 请注意:图会很大。



6.13 让每一个人用掷铜板决定先用哪只手。记录每一个人手力的差距。

6.15 (a) 画了类似图 5.2 的图。(b) 让每个人做两次, 一种温度一次, 顺序随机决定。算出每个人在不同温度下表现的差异。

6.17 (a) 根据超重多寡由小排序到大, 五个区集为: 威廉斯、邓恩、赫南德兹、摩西、圣地亚哥、肯德尔、曼、史密斯; 布伦克、欧布拉、罗德里格斯、洛伦; 杰克逊、史桃、布朗、克鲁兹; 本伯恩鲍姆、特兰、尼夫斯基、威兰斯基。(b) 最简单的方法是在每个区集内编号 1—4, 然后把区集 1 中的每个人指派到不同的减肥处理, 再把区集 2 照样做, 以此类推。

6.19 比如说所有的信封地址都用打字的, 并且用标准信封, 在同一天, 比如星期二的不同时间到同一个邮局去寄, 以便减少部分变异, 反应变量是邮递天数。有些信上写邮递区号, 有些不要写。可能的潜在变量: 目的地、寄信时是星期几, 等等。

第 7 章

7.1 可能见仁见智: (a) 似乎风险最低, 而(c) 似乎过头了。

7.9 这样是匿名, 因为名字从来不必曝光。

7.17 (a) 应该告知受试对象, 调查会问些什么问题以及大约会花多少时间。(b) 比如说, 回应者若觉得受到访员的不合理对待, 可能会想和执行机构联系。(c) 回应者不该知道出资单位(可能影响回应), 但是在公布调查结果时, 应该一并公布出资单位。

第 8 章

8.1 因为属于劳动人口的人数会有改变。

8.3 比较死亡率(校车死亡人数和私家车死亡人数分别除上各自的乘客总人数)。

8.5 7 个州当中, 德州最高(每百万人有 8.48 人被执行死刑); 佛罗里达最低(3.01)。

8.9 (a) 人口结构高龄化会使因患癌症死亡的人增加。(b) 一般人整体健康有改进, 因其他原因死亡的人数减少之后, 死于癌症的百分比会增加。(c) 若疾病能早些发现, 也就是诊断方法(而非治疗)更有效时, 存活时间就可能增长。

8.13 (a) 若长度估计是无偏的(既不偏高也不偏低), 平均应该接



近 0。如果用很多段绳子来测试,可以合理预期这样的结果。(b) 完全可靠的意思是,同条绳子重复日测的结果都一样。

8.15 (a) 偏差。(b) 比如随机指派人参加职训计划,而不是让他们自行决定参不参加。

8.17 各地警察机关也许刻意向 FBI 低报犯罪件数,或者可能在做记录方面马马虎虎。回应调查的人也许把日子记错、说谎或者不了解哪些事构成犯罪。

8.19 这个量 6 秒钟的方法,会漏掉最后一次搏动之后,到时间(6 秒钟)截止之前的“几分之一搏动”。得出的结果永远是 10 的倍数,而即使脉搏次数没变,量出来的结果也可能差 10。

8.21 需要知道班上有几人。

8.23 比较报修比例: A 牌 22%, B 牌 40%。

第 9 章

9.1 周五到周日构成一周的 42.8%。

9.3 阿纳辛和百服宁是阿司匹林的两种品牌。百服宁还包括保护胃的成分,有可能因这点而被医师指定。

9.5 比如说,20 年中有 150 000 人自杀,代表平均每天有 20 人自杀,而这种事早该引起注意。

9.7 20 项研究的 57% 等于 11.4 项研究,42% 等于 8.4 项。

9.9 (a) 应该是 210 万件,不是 2 100 万。(b) 调查中所估计的是案件(case)的计数,而不是单一事件(incident);和有暴力倾向者相处的女性(一个案件),很可能一年内被打多次(很多个事件)。

9.11 3.14%。

9.13 这样说应代表一件行李也没丢。

9.15 不应该把不同组的百分比加在一起。

9.17 要有 1 300 条横跨美国的公路才行,这实在高得离谱。

9.19 未婚女性是未婚男性的两倍多,未免差太多了。

9.21 从 1989—1997 年增加 0.64%。从 1993—1997 年增加 8.57%。

9.23 死亡事件几乎都会呈报,所以列出的死亡人数应该很准;而从 1970—1990 年增加的受伤人数,有可能是报告件数增加,而不是实际受伤人数增加。



第一部分 复习

I.3 这是方便样本。

I.5 20、11、07、24、17。

I.7 抽样误差；没有。

I.9 只有 I.8 的误差会减少。

I.11 (a) 会用网络的成人。(b) 误差界限：约 4.1%。

I.13 每位学生受访的概率是 $1/10$ ，但是每个样本必包含 3 位满 21 岁的学生及 2 位不满 21 岁的。这是分层随机样本。

I.15 用随机方式，选出 10 人用一种方法，其他人用另一种方法，再比较成功率。

I.21 可靠是指重复度量时结果差不多。要增加可靠程度可多量几次再平均。

I.23 (a) 这是观测研究，因为并没有指派处理。(b) 不像是因机遇而发生。

I.25 (a) 掉了 14.8%。

I.26 不可信：你听到一位年过 40 的女性要结婚的机会，比听说她被恐怖分子杀掉的机会大多了。可以试用《美国统计精粹》当中的资料。

第 10 章

10.1 可用饼状图。

10.3 (a) 43 148 000。(c) 可用饼状图。

10.5 这是象形图，比例有点误导。

10.7 (a) 价格波动有规则，每一年当中随供应量而上上下下。(b) 整体来说，10 年当中价格呈缓慢上升。

10.9 (a) 这是象形图，比例有点误导。

10.11 在纵轴上调整比例及最大及最小值。

10.13 (a) 下降趋势。(b) 上升趋势。(c) 无明显趋势。

10.15 可能只是季节变动(因为假日使销售额提高)。

10.17 周期约 11 年。也许有趋势，高峰点似乎在世纪中和世纪末时较高。

10.19 是。

10.21 用线图。



10.23 (a) 用线图。(b) 平均价格上升, 没有例外。(c) 最快: 1978—1980; 最慢: 1972—1974 或 1986—1987。

10.25 (a) 46%、13%、11%、4%、4% 及 22% (其他死因有 20 340 人)。(b) 用柱状图或饼状图。

10.27 (b) 冠军成绩起先不断缩短(稍有起伏), 但从 20 世纪 80 年代中期以来, 大致没什么变化。

第 11 章

11.1 大致对称, 中心点在中午附近, 散布范围从早上 6:30 到下午 5:30; 没有异常值。

11.3 强烈右偏, 在 0—4 的高峰后往右急速下降, 分布在 0—55 之间, 但没有几个大学颁发给少数族裔的工程博士超过 20 个。

11.5 茎叶图的信息会太繁杂(数字太多), 不好理解。

11.7 用直方图或茎叶图。分布大致对称, 也许有一点左偏, 分布范围从 12.7% 到 22.9%, 中心约在 17% 或 18%。

11.9 (a) 强烈右偏(许多短字, 少数很长的字)。(b) 莎士比亚用许多短字(尤其 3 个字母和 4 个字母的字), 较少很长的字。

11.11 应该会是大致对称, 可能有一点右偏。

11.13 分布的形状不规则。明显有两组不同的值, 加上左边有个值(瘦身小牛肉牌)。

11.15 大致对称, 中心在 46(“典型”的一年); 60 不是异常值。

11.17 用直方图或茎叶图。分布为右偏, 尖峰在 10—12 之间。

第 12 章

12.1 一半住户收入比这个多, 一半比这个少。

12.3 (a) 平均数较大。(b) 收入可能被夸大了。

12.5 (a) 分布大致对称。(b) $\bar{x} = 13.818\%$, $M = 13.9\%$ 。

12.7 分布右偏, 所以平均(304 万美元)比中位数(257 万美元)要大。

12.9 五数综合: 107、138.5、153、180.5、195。分布不规则; 有两组值, 中间隔开, 还有一个低异常值。

12.11 (a) 1、29、58、87、115。(b) 五数综合: 0—4, 0—4, 5—9, 10—14, 50—54。最高四分之一大约要授予至少 10 个学位。



- 12.13 分布强烈右偏，有至少两个高异常值；五数综合是 0.6、1.6、5.15、17.3、123.7。
- 12.15 有两个尖峰，所以选一个“中心”并不适当。
- 12.19 两个分布颇近似。SATM 分数通常稍高一点。
- 12.21 家禽肉热狗的卡路里含量，比其他肉类和牛肉的要低。
- 12.23 二者的平均数都是 3：(a) 分得较散， $s=2.19$ 而 (b) 的 $s=1.41$ 。
- 12.25 轿车 $\bar{x}=26.6$ ， $s=2.70$ （每加仑英里数），SUV； $\bar{x}=19.5$ ， $s=2.47$ 。
- 12.27 两组的平均数和标准差都分别是 7.50 和 2.03。A 组为左偏，而 B 组有一个高异常值。
- 12.29 这两种对离度的量度都会增加。

第 13 章

- 13.3 (a) 因为是正方形。(b) 0.5。(c) 40%。
- 13.5 84%。
- 13.7 (a) 234—298 天。(b) 少于 234 天。
- 13.9 (a) 327—345 天。(b) 16%。
- 13.11 (a) 莎拉： $z=1$ ；她的妈妈： $z=1.2$ 。(b) 莎拉的妈妈；莎拉。
- 13.13 (a) 2.5%。(b) 65—75 英寸。(c) 16%。
- 13.15 不会；它有两个尖峰，因为是由两个不同的组合并而成的（男性和女性）。
- 13.17 约 2.5%。
- 13.19 约 8%。
- 13.21 约 38%。
- 13.23 大约是第 76 百分位数。
- 13.25 约为 ± 0.7 。
- 13.27 大概要 127 分或更高。

第 14 章

- 14.1 (a) -1 到 1。(b) 任意正数。
- 14.3 (b) 约为 103 和 0.5。(c) A 和 B: GPA 很低，IQ 中等，C: 低 IQ，GPA 中等。



- 14.5 明显是正的,但不接近1。
14.7 图 14.9; 显示出较强的线性相关。
14.9 (a) 时间是解释变量。(b) 负的; 游得快比较费力。(c) 线性; 中等。
14.11 $r = 1$ 。
14.13 (a) $r = -0.746$, (b) 不会变。
14.15 因为相关关系不是线性的。
14.17 (a) 年。(b) 秒。(c) 没有单位。(d) 年。
14.19 (a) 性别并非数量变量。(b) r 不可能超过1。(c) r 没有单位。
14.21 (a) 负的。(b) 负的。(c) 正的。(d) 很小。
14.23 如果说有相关关系,也是很弱的正向相关关系。大都会和红雀队对应的点或许可视为异常值。
14.25 (a) 种植率。(c) 是曲线; 既非正相关,也非负相关。种得太挤时,收成会减低。

第 15 章

- 15.1 不爱动的孩子更有可能变胖。3.2%。
15.3 40.2%; 59.8%。
15.5 IQ 每增加1分, GPA 就上升0.101分; 8.055。
15.7 (a) 负相关,中等强度,线性或者稍微有点弯曲。
15.9 1升: 237.63; 8升: 76.84(每10万人死亡人数)。
15.11 比如说,饮食习惯和遗传(种族)背景在各国皆有不同。
15.13 (a) 负的; pH 值随时间而递减(酸度升高)。(b) 开始: 5.425; 结束: 4.635。(c) $b = -0.0053$; pH 值平均每周减低0.0053。
15.15 (a) 重量 $y = 100 + 40x$ 克; 斜率40克/周。(c) 4260克大约是9.4磅。
15.17 (a) 当 $x = 5$, $y = 5.5$; 当 $x = 10$, $y = 8$ 。(b) 只有A组。
15.19 (a) 农场人口不可能一直照图里的下降速率减少下去。
15.21 救火员的人数和造成损害的大小,是火灾严重程度的共同反应。
15.23 病的严重程度同时影响选择的医院和住院的天数。
15.25 能力强的学生比较会选这些数学课; 较差的学生可能会逃避。



- 15.27 父母不管,或者父母对学校成绩不关心。
- 15.29 有音乐素养的学生也许也有其他优势(父母较有钱、读的学校较好等)。
- 15.31 (b) 时间 $y = 43.10 - 0.0574x$, 所以预测的时间是 34.38 分钟。(c) 因为要看把哪个变量当作解释变量而有不同结果。

第 16 章

- 16.1 89.2、100、81.3。
- 16.3 (a) 降了 7.9 点(7.9%)。(b) 降了 18.7 点(18.7%)。
- 16.5 (a) 纽约的 CPI 比洛杉矶的高。(b) 我们不知道基期的物价要如何比较。
- 16.7 总价是 66.45 美元(1985 年)及 94.55 美元; GPI 是 142.3。
- 16.9 约为 38 300 美元。
- 16.11 假设 CPI 持续升高,答案应该是低于 16 美元。
- 16.13 1940 年的 32 000 美元在 1950 年约值 55 100 美元;他的真正收入增加了约 80%。
- 16.15 1976 年的 5 900 美元约等于 1999 年的 17 3000 美元;哈佛涨得比较快。
- 16.17 调整之后普度的学费为 1 158 美元、1 307 美元、1 376 美元、1 453 美元、1 490 美元、1 551 美元、1 696 美元、1 823 美元、1 889 美元、1 977 美元。普度的真正学费,在 1981—1999 年间增加了约 71%。
- 16.19 增加 3 852 美元,是 35 033 元的 11%;增加 30 324 美元,是 101 875 美元的 30%。
- 16.21 比重反映价格的差异,以及买房子和租房子的人所占百分比的差异。
- 16.23 经过季节调整的指数可能会影响某些人的收入。
- 16.25 即使在针对通货膨胀调整之后,贫困户标准仍在上升。
- 16.27 抽样愈多,信息愈多。

第二部分 复习

- II.1 有点右偏,无异常值。
- II.3 8.2%、9.3%、11.3%、14.3%、16.7%。
- II.5 11.98%。去掉密西西比州;11.79%。



- II. 7 (a) 500。(b) 68%。
- II. 9 (a) 厘米。(b) 厘米。(c) 厘米。(d) 没有单位。
- II. 11 海豚: 180 千克, 1 600 克。河马: 1 400 千克, 600 克。
- II. 13 (a) 变小。(b) 变小。
- II. 15 约 800 克。
- II. 17 接近 -1 。
- II. 19 -56.1 克; 对数据范围外做预测是很危险的。
- II. 21 约 43 990 美元。
- II. 23 不算是好的投资; 以 1983 年美元来当标准, 则在 1999 年 1 盎司黄金只值 176 美元。
- II. 25 (a) 大致对称; 最高的两个时间和最低的一个时间可视为异常值。
- II. 27 平均售价较高。
- II. 29 (a) 富达科技基金。(b) 没有。

第 17 章

17. 1 实验很多次看起来大约有 40% 出现正面。
17. 3 从表 A 得到的比例是 0.105。
17. 5 结果当然会随着图钉的种类而有所不同。
17. 7 (a) 0。(b) 1。(c) 0.01。(d) 0.6。
17. 9 (b) 个人概率可以把有关你开车习惯的特定信息都列入考虑。(c) 大部分人都自认为开车技术比一般人好。
17. 15 如果两个人谈话谈很久, 迟早会发现某个“共同点”。
17. 17 0, 0.47, 0.497, 0.4997。
17. 19 “平均数定律”用来预测天气(短期的), 并不比预测别的事可靠。
17. 23 51, 510, 5 100 和 51 000 个正面; 距离所掷总次数的一半差了 1、10、100 及 1 000 个正面。

第 18 章

18. 1 0.54。
18. 3 (a) 0.65。(b) 0.38。(c) 0.62。
18. 5 模型 1、3、4 中, 概率和不是 1; 模型 4 有超过 1 的概率。模型 2 是合法的。



- 18.7 每个可能值(1、2、3、4)的概率都是 $1/4$ 。
- 18.9 可能的点数和: 2—8; 概率 $1/16$ 、 $2/16$ 、 $3/16$ 、 $4/16$ 、 $3/16$ 、 $2/16$ 、 $1/16$ 。
- 18.11 (a) 0.1。(b) 0.3。(c) 0.5; 0.4。
- 18.13 (a) 0.438—0.502。(b) 16%。
- 18.15 0.62%。
- 18.17 (a) 约 50%。(b) 约 68%。(c) 约 32%。
- 18.19 (a) 69.4。(b) 答案随样本而不同。(c) 答案随样本而不同。

第 19 章

- 19.1 (a) 0—4 分配给民主党, 5—9 给共和党。(b) 0—5 民主党, 6—9 共和党。(c) 0—3 民主党, 4—7 共和党, 8—9 未决定。(d) 00—52 民主党, 53—99 共和党。
- 19.3 (a) 4 人选民主党, 6 人选共和党。(b) 3 人选民主党, 7 人选共和党。(c) 2 人选民主党, 4 人选共和党, 4 人未决定。(d) 6 人选民主党, 4 人选共和党。
- 19.5 (a) 0.2。(b) 0 或 1 代表 A, 2—4 代表 B, 5—7 代表 C, 8 或 9 是 D/F。
- 19.7 答案会随表 A 中起始点的选择而有不同。
- 19.9 (a) 若用 0—4 代表投中, 则他有 3 回至少中 8 次(如果 0—4 代表未中, 则有 7 回)。(b) 连续 6 次中, 连续 8 次未中(或反过来)。
- 19.11 (a) 0 或 1 代表通过, 2—9 代表不通过。(b) 0.5。(c) 不合理: 通过的概率应该会一次比一次增加。
- 19.13 (a) 0 代表窄而平的面, 1—4 代表宽而凹, 5—8 代表宽而凸, 9 代表窄而凹陷的面。(b) 结果随着表 A 起始点的选择而变。
- 19.15 (a) 0—8 代表系统 A 正常运作。(b) 0—7 代表系统 B 正常运作。(c) 结果随表 A 的起始点而变。
- 19.17 00—24 代表乘客未出现。结果随表 A 的起始点而变。
- 19.19 0—5 代表备用车可到机场载客。
- 19.21 (a) 男男男, 男男女, 男女男, 女男男, 女女男, 女男女, 男女女, 女女女。第一项的概率是 0.132 651; 其次三项 0.127 449; 再接着三项 0.122 451; 最后一项 0.117 649。(实际应用时, 会把这些数字四舍五入到小数点后 2 或 3 位。)(b) 0.867 349。



第 20 章

- 20.1 0.60 美元。
- 20.3 约 0.947 4 美元。
- 20.5 约 0.947 4 美元(和 20.3 题一样)。
- 20.7 (a) 第一个序列。(b) 第二个序列看起来“比较随机”。
- 20.9 (a) 0.75 美元。(b) 0.72 美元。
- 20.11 (a) 1.3 个雌性后代。(b) (平均来说)每一代的雌虫是上一代的 1.3 倍。
- 20.13 0—1 代表 0 个雌性后代, 2—4 代表 1 个, 5—9 代表 2 个。结果因表 A 始点而变。
- 20.15 2.55 个人。
- 20.17 00—48 代表女孩, 49—99 代表男孩。结果会随表 A 的起始点而变。
- 20.19 0—1 代表猜对, 2—9 代表猜错。结果会随表 A 的起始点而变。
- 20.21 $np = 0.5$; 模拟结果为 0.502。

第三部分 复习

- Ⅲ.1 (b) 如果 10 个数字的概率一样, 应该是 30%。
- Ⅲ.3 (a) 0.1。(b) 用 0—3 当 O 型, 4—6 当 A 型, 7 和 8 当 B 型, 9 当 AB 型。
- Ⅲ.5 结果会随表 A 的起始点而变。
- Ⅲ.7 3.5。
- Ⅲ.9 (b) 0.03。(b) 00—91 是一对或更差, 92—99 是两对或更好。
- Ⅲ.11 (a) 所有概率均在 0 和 1 之间, 总和为 1。(b) 0.41。(c) 0.38。
- Ⅲ.13 (a) 0.27。(b) 0.57。
- Ⅲ.15 (a) 68%。(b) 95%。
- Ⅲ.17 0.70%。
- Ⅲ.19 该有的概率: 每个“面值”概率 $1/13$ 。
- Ⅲ.21 密歇根、俄亥俄州立大学及普度: 0.155 6(经过四舍五入); 伊利诺伊、印第安纳和威斯康星: 0.077 8(也经过四舍五入)。



第 21 章

- 21.1 (a) 唐雅宿舍的 175 位住宿同学。(b) 喜欢宿舍所提供食物者的比例。(c) $\hat{p} = 0.28$ 。
- 21.3 产生这个区间所用的方法, 有 95% 时候会得出正确结果。
- 21.5 (a) 所有女性中认为个人时间不够者所占比例。(b) 0.439 到 0.501。
- 21.7 (a) 0.564 — 0.636。(b) 0.575 — 0.625。(c) 0.582 — 0.618。(d) 样本大小增加时, 区间会变窄。
- 21.9 不同的图钉会有不同的结果。
- 21.11 0.491—0.523; 0.5 在此区间中, 所以正面概率可能是 1/2。
- 21.13 (a) 平均数为 0.14, 标准差为 0.0142 的正态分布。(b) 至少 18.2% 的机会不大 (0.15%); 至少 11.2% 则很可能 (97.5%)。
- 21.15 速算法: 0.0447; 新方法: 0.307。
- 21.17 (a) 用 0—5 代表支持柯可丝者。(b) 结果会随表 A 的起始点而变。
- 21.19 0.43—0.51; 此区间较宽。
- 21.21 0.0665—0.1026。
- 21.23 当 $\hat{p} = 0.5$ 时, $\hat{p}(1 - \hat{p}) = 0.25$; 其他值都比这个小。

第 22 章

- 22.1 如果上教堂的人, 种族优越感不比不上教堂的人高, 那么像我们观测到的这种样本结果将会很难得发生。
- 22.3 因为 $P = 0.02$ 代表, 降 6.7% 是因碰巧产生的概率只有 2%。
- 22.5 如果猪头骨没有特别意义, 则所观察到的差异其发生机会小于 1%。
- 22.7 (a) 这是方便样本。(b) 这里的“显著超过”代表超过不少。(c) 如果大家不想要更多钱的话, 所观察到的结果发生的机会将小于 1%。
- 22.9 (a) 体温低于 98.6°F 者所占比例。(b) $H_0: p = 0.5$; $H_a: P > 0.5$ 。
- 22.11 $H_0: P = 0.21$; $H_a: P \neq 0.21$ 。
- 22.13 (a) 这所学校所有一年级学生中, 想要有钱的比例。(b) $H_0: p = 0.73$; $H_a: p \neq 0.73$ 。(c) 0.66; $\hat{p} < 0.66$ 或 $\hat{p} > 0.79$ 的这个事件。(d) 若 H_0 正确, 像 0.66 这样或更极端的 \hat{p} 值, 发生的概率只有 3.7%。



- 22.15 有 5% 的统计显著性水平, 但是没有 1% 的统计显著性水平。
- 22.17 不对; 这是说如果 H_0 为真, 则我们所观察到的结果, 发生的机会小于 5%。
- 22.19 (a) $H_0: p=0.2$; $H_a: p>0.2$ 。(b) 用 0 或 1 当作猜中, 2—9 当作没猜中。(c) 0 个对有 1 次, 1 个对有 4 次, 2 个对有 9 次, 3 个对有 5 次, 6 个对有 1 次。(d) $\hat{p}>0.5$ 的这个事件。我们估计 $P=0.05$ 。
- 22.21 标准计分 -2.23; 没有 $\alpha=0.025$ 的统计显著性。

第 23 章

- 23.1 我们的置信区间公式, 只能用在 SRS 上。
- 23.3 我们的置信区间中包含代表乙会赢的 p 值, 而真正的估计误差还可能更大。
- 23.5 (a) 大小为 500 的样本中, 由于 $P<0.01$ 我们预期会看到大约 5 个人有超能力。(b) 再测试一下这 4 个人。
- 23.7 是否是随机样本? 样本多大?
- 23.9 基本上正确。
- 23.11 (a) 5%。(b) 有些会因机遇而显示出具统计显著性的差异。
- 23.13 (a) 选 50 人给他们疫苗; 比较受感染的比例。(b) $H_0: P_1=P_2$; $H_a: P_1>P_2$ 。(c) 所观察到的这种差异, 光因为机遇就会有 25% 时候发生。(d) 同意。
- 23.15 0.594—0.726。
- 23.17 906 人通过: 0.888—0.924。907 人通过: 0.889—0.925。

第 24 章

- 24.1 各组成绩 C 或更佳所占百分比分别为 55%, 74.7%, 37.5%。花一点(但不能太多)时间在课外活动似乎有好处。
- 24.3 (a) 孵出: 16、38、75; 未孵出: 11、18、29。(b) 冷: 59.3%; 中等: 67.9%; 暖: 72.1%。冷水并不会使蛋孵不出, 只是使孵出的机会减低。
- 24.5 (a) 693 000; 表中数字四舍五入到千位。(b) 55.1%、55.9%、41.6%、40.0%。女性得到大部分的学士和硕士, 但是专业学位和博士学位的女性百分比就较低。
- 24.7 先把 a 定成是任何 10—50 之间的数。
- 24.9 (a) 白人被告: 19 个死刑, 141 个非死刑。黑人被告: 17 个



死刑, 149 个非死刑。(b) 整体死刑比例: 白人被告 11.9%, 黑人被告 10.2%。被害人为白人, 则白人和黑人被告被判死刑的比例分别为 12.6% 及 17.5%; 而被害人为黑人时, 则分别为 0% 及 5.8%。(c) 死刑在被害人是白人时(14%)比被害人为黑人(5.4%)时容易发生。白人被告 94.3% 时候杀的是白人, 但是和杀白人的黑人比起来, 被判死刑的机会较小。

24.11 (a) 列总和: 82、37。行总和: 20、91、8。全部总和: 119。预期计数: 13.8、62.7、5.5; 6.2、28.3、2.5。中间一行的计数差最多。(b) 6.93; 右下那格的“贡献”最大, 中间那行稍小。(c) $df = 2$, 有 $\alpha = 0.05$ 的统计显著性水平, $\alpha = 0.01$ 时则无。

24.13 $\chi^2 = 1.703$, $df = 2$, 无统计显著性。

24.15 (a) $\chi^2 = 14.863$; $P < 0.002$ 。差异具统计显著性。(b) 每个城市的数据应来自(不受规范的)托儿保姆的 SRS, 不同城市的样本互相独立。

第 25 章

25.3 (a) 够接近正态分布了, 不过有两个低异常值。 $\bar{x} = 105.84$, $s = 14.27$ 。(b) 100.8—110.9。(c) 我们的置信区间公式, 是在数字代表 SRS 时才适用。

25.5 标准计分 -2.28; 有 $\alpha = 0.025$ 的统计显著性水平。

25.7 (a) 平均数为 115、标准差为 6 的正态分布。(b) 118.6 相当靠近曲线的中间位置, 如果 H_0 为真, 这种结果不令人意外。而 125.7 落在右边尾巴的高处, 当 $\mu = 115$ 时, 这种结果很少出现。

25.9 $H_0: \mu = 18$, $H_a: \mu < 18$ 秒。

25.11 标准计分 1.01; 不具统计显著性。

25.13 0.480—0.584。

25.15 勉强有正态分布的形态。

25.17 (a) $\bar{x} = 0.496$, 离 0.5 很近。(b) 估计 $\sigma = 0.3251$ 。

第四部分 复习

IV.1 0.669—0.750。

IV.3 约 66%—74%。

IV.5 0.208—0.272。

IV.7 你有所有州的信息, 而不是只有样本。



IV.9 (a) 如果两组医师之间没有区别, 像这样的结果难得发生。

(b) 有两个比例可以拿来比较。

IV.11 上午: 0.550—0.590。晚间: 0.733—0.767。晚间的比例高了不少。

IV.13 (a) $H_0: p = 0.57$; $H_A: p > 0.57$ 。(b) 平均数为 0.57、标准差为 0.01 的正态分布。(c) 是。

IV.15 (a) 有些人不愿承认没有定期上教堂, 或者以为自己常常去, 实际上却没去那么多。(b) 样本结果本就和总体真正值有异。

(c) 得到两个区间所用的方法, 都是在 95% 的时候正确。

IV.17 $H_0: p = 1/3$; $H_A: p > 1/3$; 标准计分 1.82; 有 $\alpha \approx 0.05$ 的统计显著性水平而没有 $\alpha = 0.025$ 的统计显著性水平。

IV.19 0.247—0.313。

IV.21 $H_0: p = 0.57$; $H_A: p > 0.57$; 标准计分 17.99; 对应任何合理的 α 值都有统计显著性。

IV.23 标准计分 -2.77; 有 $\alpha \approx 0.005$ 的统计显著性。

IV.25 (b) $\bar{x} \approx 15.59$ 英尺, $s = 2.550$ 英尺; 15.0—16.2 英尺。(c) 我们在考虑怎样的总体? 成年鲨鱼? 雄鲨?

IV.27 标准计分 1.53; 有 $\alpha = 0.10$ 的统计显著性水平, 没有 $\alpha = 0.05$ 的统计显著性水平。

IV.29 (a) 2.96%、13.07%、6.36%。(b) 两列分别为 721、173、412; 22、26、28。(c) 预期计数: 702、14、188.06、415.80; 40.86、10.94、24.20。所有预期计数都大于 5。(d) H_0 : 是否申诉和是否退出 HMO 没有相关关系; H_A : 二者之间有相关关系。 $df = 2$; 非常有统计显著性。

IV.31 (a) 列分别为 1313、1840; 991、614。上午: 57.0%, 晚间: 75.0%。(b) 我们的样本很大, 而两项比例差很多。(c) $\chi^2 = 172$, $df = 1$; 非常有统计显著性。

表 A 随机数字

列

101	19223	90534	05756	28713	96409	12531	42544	82853
102	73676	47150	99400	01927	27754	42648	82425	36290
103	45467	71709	77558	00095	32863	29485	82226	90056
104	52711	38889	93074	60227	40011	85848	48767	52573
105	95592	94007	69971	91481	60779	53791	17297	59335
106	68417	35013	15529	72765	85089	57067	50211	47487
107	82739	57890	20807	47511	81676	55300	94383	14893
108	60940	72024	17868	24943	61790	90656	87964	18883
109	36009	19365	15412	39638	85453	46816	83485	41979
110	38448	48789	18338	24697	39364	42006	76688	08708
111	81486	69487	60513	09297	00412	71238	27649	39950
112	59639	88804	04634	71197	19352	73089	84898	45785
113	62568	70206	40325	03699	71080	22553	11486	11776
114	45149	32992	75730	66280	03819	56202	02938	70915
115	61041	77684	94322	24709	73698	14526	31893	32592
116	14459	26056	31424	80371	65103	62253	50490	61181
117	38167	98532	62183	70632	23417	26185	41148	75532
118	73190	32533	04470	29669	84407	90785	65956	86382
119	95857	07118	87664	92099	58806	66979	98624	84826
120	35476	55972	39421	65850	04266	35435	43742	11937
121	71487	09984	29077	14863	61683	47052	62224	51025
122	13873	81598	95052	90908	73592	75186	87136	95761
123	54580	81507	27102	56027	55892	33063	41842	81868
124	71035	09001	43367	49497	72719	96758	27611	91596
125	96746	12149	37823	71868	18442	35119	62103	39244
126	96927	19931	36809	74192	77567	88741	48409	41903
127	43909	99477	25330	64359	40085	16925	85117	36071
128	15689	14227	06565	14374	13352	49367	81982	87209
129	36759	58984	68288	22913	18638	54303	00795	08727
130	69051	64817	87174	09517	84534	06489	87201	97245
131	05007	16632	81194	14873	04197	85576	45195	96565
132	68732	55259	84292	08796	43165	93739	31685	97150
133	45740	41807	65561	33302	07051	93623	18132	09547



(续表)

134	27816	78416	18329	21137	35213	37741	04312	68508
135	66925	55658	39100	78458	11206	19876	87151	31260
136	08421	44753	77377	28744	75592	08563	79140	92454
137	53645	66812	61421	47836	12609	15373	98481	14592
138	66831	68908	40772	21558	47781	33586	79177	06928
139	55588	99404	70708	41098	43563	56934	48394	51719
140	12975	13258	13048	45144	72321	81940	00360	02428
141	96767	35964	23822	96012	94591	65194	50842	53372
142	72829	50232	97892	63408	77919	44575	24870	04178
143	88565	42628	17797	49376	61762	16953	88604	12724
144	62964	88145	83083	69453	46109	59505	69680	00900
145	19687	12633	57857	95806	09931	02150	43163	58636
146	37609	59057	66967	83401	60705	02384	90597	93600
147	54973	86278	88737	74351	47500	84552	19909	67181
148	00694	05977	19664	65441	20903	62371	22725	53340
149	71546	05233	53946	68743	72460	27601	45403	88692
150	07511	88915	41267	16853	84569	79367	32337	03316

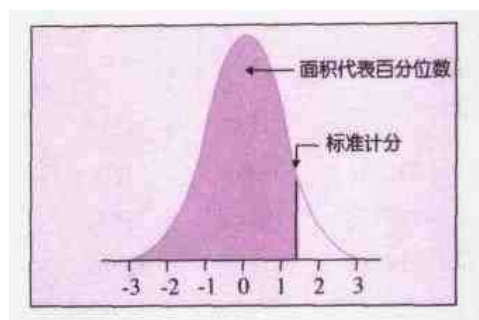


表 B 正态分布的百分位数

标准计分	百分位数	标准计分	百分位数	标准计分	百分位数
-3.4	0.03	-1.1	13.57	1.2	88.49
-3.3	0.05	-1.0	15.87	1.3	90.32
-3.2	0.07	-0.9	18.41	1.4	91.92
-3.1	0.10	-0.8	21.19	1.5	93.32
-3.0	0.13	-0.7	24.20	1.6	94.52
-2.9	0.19	-0.6	27.42	1.7	95.54
-2.8	0.26	-0.5	30.85	1.8	96.41
-2.7	0.35	-0.4	34.46	1.9	97.13
-2.6	0.47	-0.3	38.21	2.0	97.73
-2.5	0.62	-0.2	42.07	2.1	98.21
-2.4	0.82	-0.1	46.02	2.2	98.61
-2.3	1.07	0.0	50.00	2.3	98.93
-2.2	1.39	0.1	53.98	2.4	99.18
-2.1	1.79	0.2	57.93	2.5	99.38
-2.0	2.27	0.3	61.97	2.6	99.53
-1.9	2.87	0.4	65.54	2.7	99.65
-1.8	3.59	0.5	69.15	2.8	99.74
-1.7	4.46	0.6	72.58	2.9	99.81
-1.6	5.48	0.7	75.80	3.0	99.87
-1.5	6.68	0.8	78.81	3.1	99.90
-1.4	8.08	0.9	81.59	3.2	99.93
-1.3	9.68	1.0	84.13	3.3	99.95
-1.2	11.51	1.1	86.43	3.4	99.97

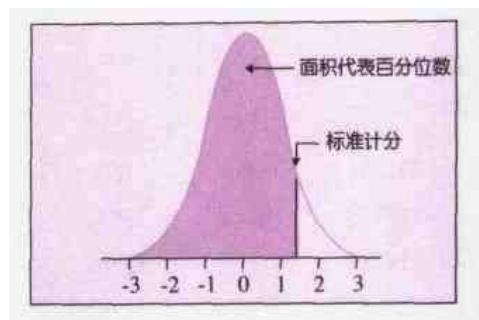


表 B 正态分布的百分位数

标准计分	百分位数	标准计分	百分位数	标准计分	百分位数
-3.4	0.03	-1.1	13.57	1.2	88.49
-3.3	0.05	-1.0	15.87	1.3	90.32
-3.2	0.07	-0.9	18.41	1.4	91.92
-3.1	0.10	-0.8	21.19	1.5	93.32
-3.0	0.13	-0.7	24.20	1.6	94.52
-2.9	0.19	-0.6	27.42	1.7	95.54
-2.8	0.26	-0.5	30.85	1.8	96.41
-2.7	0.35	-0.4	34.46	1.9	97.13
-2.6	0.47	-0.3	38.21	2.0	97.73
-2.5	0.62	-0.2	42.07	2.1	98.21
-2.4	0.82	-0.1	46.02	2.2	98.61
-2.3	1.07	0.0	50.00	2.3	98.93
-2.2	1.39	0.1	53.98	2.4	99.18
-2.1	1.79	0.2	57.93	2.5	99.38
-2.0	2.27	0.3	61.97	2.6	99.53
-1.9	2.87	0.4	65.54	2.7	99.65
-1.8	3.59	0.5	69.15	2.8	99.74
-1.7	4.46	0.6	72.58	2.9	99.81
-1.6	5.48	0.7	75.80	3.0	99.87
-1.5	6.68	0.8	78.81	3.1	99.90
-1.4	8.08	0.9	81.59	3.2	99.93
-1.3	9.68	1.0	84.13	3.3	99.95
-1.2	11.51	1.1	86.43	3.4	99.97

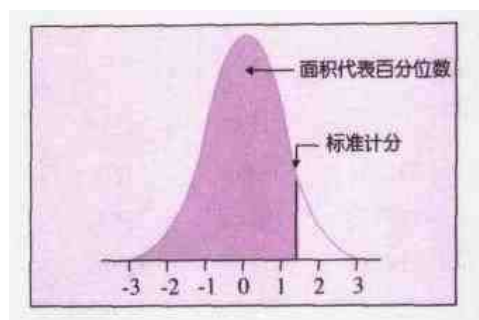


表 B 正态分布的百分位数

标准计分	百分位数	标准计分	百分位数	标准计分	百分位数
-3.4	0.03	-1.1	13.57	1.2	88.49
-3.3	0.05	-1.0	15.87	1.3	90.32
-3.2	0.07	-0.9	18.41	1.4	91.92
-3.1	0.10	-0.8	21.19	1.5	93.32
-3.0	0.13	-0.7	24.20	1.6	94.52
-2.9	0.19	-0.6	27.42	1.7	95.54
-2.8	0.26	-0.5	30.85	1.8	96.41
-2.7	0.35	-0.4	34.46	1.9	97.13
-2.6	0.47	-0.3	38.21	2.0	97.73
-2.5	0.62	-0.2	42.07	2.1	98.21
-2.4	0.82	-0.1	46.02	2.2	98.61
-2.3	1.07	0.0	50.00	2.3	98.93
-2.2	1.39	0.1	53.98	2.4	99.18
-2.1	1.79	0.2	57.93	2.5	99.38
-2.0	2.27	0.3	61.97	2.6	99.53
-1.9	2.87	0.4	65.54	2.7	99.65
-1.8	3.59	0.5	69.15	2.8	99.74
-1.7	4.46	0.6	72.58	2.9	99.81
-1.6	5.48	0.7	75.80	3.0	99.87
-1.5	6.68	0.8	78.81	3.1	99.90
-1.4	8.08	0.9	81.59	3.2	99.93
-1.3	9.68	1.0	84.13	3.3	99.95
-1.2	11.51	1.1	86.43	3.4	99.97

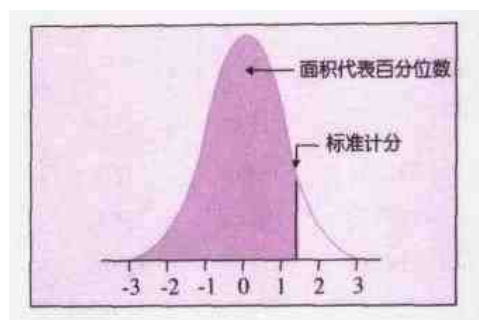


表 B 正态分布的百分位数

标准计分	百分位数	标准计分	百分位数	标准计分	百分位数
-3.4	0.03	-1.1	13.57	1.2	88.49
-3.3	0.05	-1.0	15.87	1.3	90.32
-3.2	0.07	-0.9	18.41	1.4	91.92
-3.1	0.10	-0.8	21.19	1.5	93.32
-3.0	0.13	-0.7	24.20	1.6	94.52
-2.9	0.19	-0.6	27.42	1.7	95.54
-2.8	0.26	-0.5	30.85	1.8	96.41
-2.7	0.35	-0.4	34.46	1.9	97.13
-2.6	0.47	-0.3	38.21	2.0	97.73
-2.5	0.62	-0.2	42.07	2.1	98.21
-2.4	0.82	-0.1	46.02	2.2	98.61
-2.3	1.07	0.0	50.00	2.3	98.93
-2.2	1.39	0.1	53.98	2.4	99.18
-2.1	1.79	0.2	57.93	2.5	99.38
-2.0	2.27	0.3	61.97	2.6	99.53
-1.9	2.87	0.4	65.54	2.7	99.65
-1.8	3.59	0.5	69.15	2.8	99.74
-1.7	4.46	0.6	72.58	2.9	99.81
-1.6	5.48	0.7	75.80	3.0	99.87
-1.5	6.68	0.8	78.81	3.1	99.90
-1.4	8.08	0.9	81.59	3.2	99.93
-1.3	9.68	1.0	84.13	3.3	99.95
-1.2	11.51	1.1	86.43	3.4	99.97